Age-Related Differences in Object-Similarity Judgment

by

D Estes M^{c}Knight

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science
University of Alberta

© D Estes M^{c}Knight, 2022

Abstract

As people age, they adjust how they associate objects. We examine the interplay between age and various aspects of object similarity. For this, we perform two experiments: one between adults aged 25–35 and 50–60 and another between six-year-olds and adults. Between 25–35- and 50–60-year-olds, we investigate discrepancies in preference for each of 49 object-comparison dimensions. Between six-year-olds and adults, we investigate changes in the prioritization of the classes of object-comparison relations known as taxonomic and thematic relations. To facilitate these tasks, we use a prior interpretable, machine-learned computational model; this model is trained to perform an odd-one-out-among-three task with a vector embedding for each object being compared.

For the first task, between 25–35-year-olds and 50–60-year-olds, we examine each of the 49 object-comparison dimensions defined by the prior interpretable model. These dimensions are human-interpretable, quantifying similarities such as "metallic" or "food-related." We modify the architecture of that model to learn the preferences of each age group for each of those dimensions. We then compare those preferences between each age group.

For the second task, between six-year-olds and adults, we examine the classes of object-comparison relations described by taxonomic and thematic reasoning. We use the interpretable model to assign taxonomic and thematic scores to object-comparison questions, then select questions from amongst those to administer to six-year-olds.

Finally, we contrast their responses with previous adult ones to elicit age-related changes in the prioritization of taxonomic and thematic reasoning, both in absolute terms and relative to one another.

In the context of prior literature, we provide measures of differences in object-similarity judgment between younger and older adults for each of 49 fine-grained types of object similarity and remark upon the resulting trends. We corroborate a thematic-to-taxonomic trend in thinking from adolescence to adulthood. Finally, we expand the knowledge-base of the common-resource THINGS initiative with our results.

Preface

This thesis is an original work by D Estes McKnight under the supervision of Professor Alona Fyshe at the University of Alberta.

The portions of this thesis involving child-response data are part of a research project that received ethics approval from the Conjoint Faculties Research Ethics Board (CFREB) at the University of Calgary under the title, “Do Young Children and Adults Categorize Images of Objects Similarly? A Comparison Study Using Images from the THINGS Database in an Odd-One-Out Paradigm” (No. REB20-2002, February 17, 2021). These parts were additionally done in collaboration with Alona Fyshe, Suzanne Curtin, Kelly Burkinshaw, Janet Werker, Henny Yeung, Patrick Mihalicz, and Alexis Black. Child-response data were collected by Patrick Mihalicz and Kelly Burkinshaw at the Speech Development Lab under Suzanne Curtin at the University of Calgary.

Animacy-determination work used in this paper was previously presented as “Using Wiktionary and Automatic Translation for Object-Animacy Annotation” online at the 2022 Hidden Methods conference on April 27, 2022. No other part of this thesis has been previously published. All parts of this thesis may, in the future, be submitted for publication.

"Why is a raven like a writing-desk?" riddled the Hatter.

"I'm not sure. In what manner?" asked Alice.

"I haven't the slightest idea."

—Lewis Carroll*, *Alice's Adventures in Wonderland*

(*with some creative liberty for brevity)

Acknowledgements

To my supervisor, Alona Fyshe, for taking me on and supporting me as her graduate student. You have given me the invaluable opportunities to teach and research in the course of my program, for which I am extremely grateful.

To the other members of my defence committee, Carrie Demmans Epp, Denilson Barbosa, and Henny Yeung, for giving your time to review this thesis and advance me on this academic journey.

To Suzanne Curtin, Kelly Burkinshaw, Janet Werker, Henny Yeung, Patrick Mihalicz, and Alexis Black, whom I had the pleasure of working with and learning from in investigating children's behavior.

To Antonie Bodley. Without the regular meetings with you on this thesis, the finishing thereof would have been far less timely.

Finally, to my parents, William and Karen McKnight, for your continued advocacy of me regardless of identification or quirk, and for steadfastly guiding me again and again through challenges experienced and incurred. Your unwavering support means the world.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

In our interactions with our environments, we conceptualize extraordinarily diverse surroundings, situating and reasoning with the objects within them irrespective of our familiarity. One way we achieve this is by interpreting the objects in the context of those we are more familiar with; that is, judging their similarity. Object similarity informs how we act upon objects [1], assists in the recall of information, and shapes how we interpret unfamiliar entities from a young age [2]. Despite consequent interest in the topic, however, uncovering the specifics of how we judge the similarity between objects is a many-layered task. This is likewise true for how this ability develops as we age. In this thesis, our research goal is using new advances in modelling object similarity to examine how people evolve their means of object comparison with age.

1.1 Overview

Human judgment of object similarity is not a static procedure. Not only do different individuals diverge in performing object comparison, but a given individual's perception varies based on the presentation and reasons for comparison [3]. In spite of these differences, people holistically exhibit enough consistency that population-level trends emerge, such as with image similarity [4] and object styles [5]. The level at which people diverge from one another can vary considerably with the group and axis of similarity in question; i.e., with various language features [6, Introduction] and with expertise and animal categorization [7]. One area of particular interest in this regard is the impact of age on object-similarity judgments.

The task of object comparison involves many factors, and accordingly, prior work abstracts these factors into broader classes for evaluation. Some of these generalizations include objects being related due to shared observable features or functionality, co-occurring in the real world, word co-occurrence ("cottage cheese"), hypernymy, and combinations thereof [8]. For example, a relationship between objects along their shared features, such as a dog and a bear both having paws and fur, is classified as a taxonomic relation. Meanwhile, a relationship between two objects due to them co-occurring, such as between a dog and a leash, is referred to as a thematic relation. Broad classes of relations such as these have received extensive attention, particularly concerning adolescent development, and prior work on taxonomic and thematic relations specifically has found age-related differences in people's prioritization thereof [9], [10].

There are a number of ways to model these similarities, but two broad approaches [11] are as follows. The first is representing them with a set of dimensions, then using these dimensions to compute their similarity. The second is modelling them as similar by virtue of them exhibiting shared sets of (binary) present-or-not-present properties [12]. There is some overlap between these approaches, as a dimension can continuously scale between binary property values. In this thesis, we use a recent method from Hebart et al. [13] that combines traits of both of these approaches in this manner into a single model. Their approach produces a computational model with nonnegative dimensions identifiable as object-similarity properties. This model is produced from responses to an object-comparison task on a set of objects [14] that form the core of the THINGS initiative [15], an effort to link various research results to the same shared objects.

We use the Hebart et al. model and its properties to answer two questions about the relationship between age and object similarity for the THINGS objects:

1. For our first task, we examine differences between younger (ages 25–35) and older (ages 50–60) adults in judging object similarity on a set of narrow types of object similarity. To do this, we consider the Hebart et al. similarity model, whose dimensions have human-identifiable labels. We construe the usages of the Hebart et al. model's dimensions as signifying usages of latent individual types of similarities described by these human labels. We modify the model to learn preferences for these types of similarities for 25–35-year-olds and 50–60-year-olds. We then compare the preferences between the two age groups.

2. For our second task, we examine differences between adults (age 18 and up) and young children (age 6) in judging broad types of object similarity. To do this, we construct approximations of taxonomic- and thematic-relation usage from the model's dimensions. Then, we select sets of three objects to administer to children as part of an object-comparison task based on these approximations. We record the children's responses and compare them with prior adult responses to determine differences between the age groups in usage.

1.2 Contributions and Thesis Layout

1.2.1 Contributions

Considerable work has been done on the relationship between age and object comparison. Here, we address the motivation for each of our thesis questions independently. We also summarize this thesis's contributions to the literature.

Our first task examines differences between younger (ages 25–35) and older (ages 50–60) adults when comparing objects using narrow types of object similarity. Evaluating differences in how adults of different ages perform object comparisons has been the focus of past work; however, that research has typically concentrated on broad classes of relations, such as taxonomic and thematic relations, rather than on more fine-grained properties (such as “metallic”). Additionally, to our knowledge, no prior work has shown these associations for similarity-task-derived features representative of human-identifiable object-relational properties.

Our second task examines differences between children (age 6) and older (ages 18–110) adults in comparing objects along broad taxonomic and thematic lines. Prior research on this has looked at the usages of taxonomic and thematic relations relative to each other, rather than independently.

Beyond all of these prior-body-of-work factors, knowing how age interacts with the 49 similarity dimensions computed for the THINGS collaborative-research initiative contributes to that initiative.

In light of the above motivation, this thesis’s primary contributions are as follows:

1. We provide measures of the differences between 25–35-year-olds and 50–60-year-olds in their preferences for specific types of comparison, namely those described by the Hebart et al. model’s dimensions. We also determine for which of these relatively fine-grained types of object similarity the age groups differ significantly. We discuss the trend revealed by the largest of these differences.

2. We provide a framework for determining differences between six-year-olds and adults in their usage of taxonomic and thematic relations, where these taxonomic and thematic preferences are not relative to each other. We also weakly corroborate prior work showing a taxonomic-to-thematic trend in adolescent development.

3. With both groups of contributions, we expand the THINGS initiative’s common pool of knowledge.

1.2.2 Thesis Layout

This thesis addresses two major topics concerning object-similarity judgment and age: one focused on younger and older adults, and another focused on young children and adults of general age. To address this, the project for each topic is self-contained, with the between-adult-ages topic being the target of Chapter 3 and the child-vs-adult topic being the topic of Chapter 4. Information common to each can be found in the background chapter, Chapter 2, although some background information specific to each chapter in isolation is contained in that chapter's own constituent introduction section. Each project chapter also contains its own methodology, results, and discussion sections. Finally, Chapter 5 summarizes this thesis's major findings.

Chapter 2

Background

In this chapter, I give context to my thesis, specifically focusing on the following areas:

- Object similarity and representation thereof
- Age and taxonomic/thematic similarity
- Neural networks
- The Hebart et al. model

2.1 Object Similarity and its Representation

People routinely perform comparisons to contextualize concepts and interpret their place in the world. Dogs and bears are similar because they are mammals, sadness and the state of being upset are similar in terms of being negative emotions, and shampoo and showers are similar because they are associated with bathing (and, to an English speaker, perhaps because the label for each concept begins with the same sound). In this manner, concepts interact with each other in myriad ways (features, co-occurrence, related concepts), and at different levels (being mammals vs being animals, co-occurring at Lake Michigan vs. co-occurring at the Great Lakes).

Non-exhaustively, some prominent types of similarity include the following:

- Function-based similarity, wherein objects perform similar functions (such as brooms and vacuums both cleaning floors)
- Structural similarity [16], wherein objects share relationships between elements (such as between an atom and the Solar System: electrons orbit the nucleus within an atom and the planets orbit the Sun within the Solar System)
- Thematic similarity, wherein objects appear in similar contexts (such as dogs and bones both appearing in a pet context)
- Lexical coöccurrence-based similarity, wherein objects' associated words co-occur (such as cottage and cheese both appearing in the phrase cottage cheese)
- Various lexical semantic similarities, such as hyponymy/hypernymy (dog-mammal) and meronymy/holonymy (dog-paw)
- Category-based similarity, wherein two objects belong to some abstract category (such as "is an Apple computer"). This is further broken down into "basic categories," or categories more fundamental for comparison [17], and "ad-hoc categories," or categories constructed for some arbitrary purpose [3]. The distinction between these is whether the category exists in memory outside of that purpose.

These similarities have some overlap. For example, when defining a category-based similarity of "has fur," this correlates with two given objects having paws, which in turn correlates with the two objects being mammals and sharing mammalian structural similarities. Defining a categorical or structural similarity of "objects with four legs and a cushion" will correspond with many objects being functionally related by virtue of being "able to be sat upon."

Similarities can be encoded in a number of ways, including as relational graphs (i.e., a hypernymy structure imposed by WordNet [18]), objects as sets of binary features (i.e., Tversky's contrast model [12]), and objects as vectors of continuous features (i.e., word embeddings [19]). The choice of operation on these representations to produce quantifications of similarity can also vary considerably; three such for continuous features include dot products, cosine similarity, and distance metrics.

2.2 Age and Taxonomic/Thematic Similarity

Two prominent ways of comparing objects are known as taxonomic and thematic relations [8]. Taxonomic relations are relations where objects are grouped by shared features (dogs and bears have four legs, are furry, etc) whereas thematic relations are those where objects are contextually related (dogs and bones, bears and trees). Objects can be related by both taxonomic and thematic means (dogs and cats). A visual depiction of these is given in Figure 2.1.

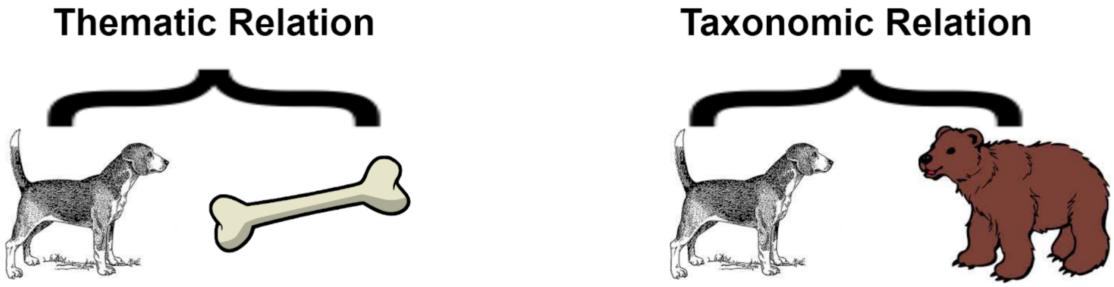
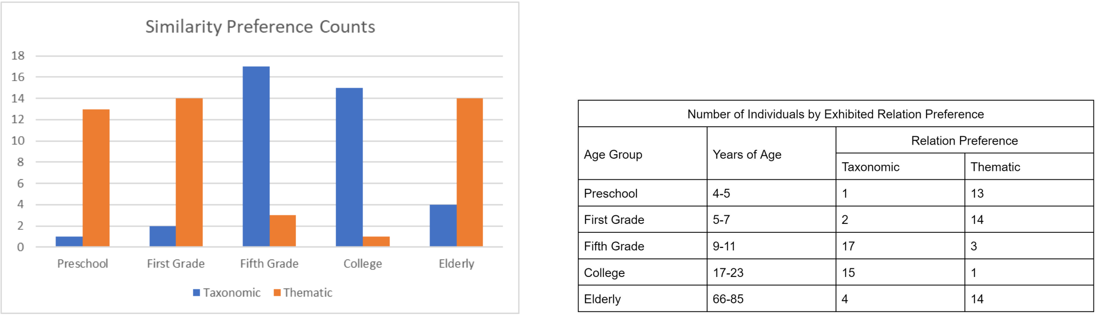


Figure 2.1: Thematic and taxonomic relation illustration.

In the literature, the usage of taxonomic can differ somewhat, but here we define it to include category-based, function-based, and structure-based similarities. Beyond definitional differences, there are observable real-world differences between the two types of relations, such as in terms of processing [20], situational usage, and development. Importantly, the relative usages of both of these relations change with age.

From a young age, children have an awareness of how to apply both taxonomic and thematic relations [21], [22]. Young children apply taxonomic relations when primed to identify objects with a new, unknown lexical label, for example. A common task to evaluate children's usage of taxonomic and thematic relations is the ***triad task***, wherein a participant is given an initial object, then presented with two additional objects and asked to identify the most similar. In one study that used the triad task with taxonomically similar and thematically similar objects, giving an unknown label to the initial object prompted an elevated level of taxonomic-relationship prioritization in children [2].

Another well-known study used the triad task to evaluate taxonomic and thematic preferences of different age groups [9]. They found that respondents of different ages exhibited different preferences for taxonomic and thematic relations, with an initial preference for thematic relationships in first grade giving way to a taxonomic preference by fifth grade, then reverting again to a thematic preference sometime between the average college age and old age. These specifics are given in Figure 2.2.



Number of Individuals by Exhibited Relation Preference

Age Group	Years of Age	Relation Preference	
		Taxonomic	Thematic
Preschool	4-5	1	13
First Grade	5-7	2	14
Fifth Grade	9-11	17	3
College	17-23	15	1
Elderly	66-85	4	14

Figure 2.2: Relative taxonomic and thematic relation preference with age. Results from Smiley and Brown [9].

This earlier thematic-preference-to-taxonomic shift has been observed in children as young as two to three years of age. One study [22] used a match-to-sample task[1] with positive reinforcement for the identification of two highly similar objects along taxonomic or thematic grounds to test the relative taxonomic and thematic preferences of young children (aged 2–3). Children aged 26 months identified 46.5% of thematic similarities and between 52-83% of all but the most coarse-grained taxonomic similarities, while children at 3 years of age identified a much higher 65.8% of thematic similarities

[1] A task wherein children are given some base object/concept and asked to identify a match from a set of options

and between 66-87% of all but the most coarse-grained taxonomic similarities.

2.3 Neural Networks

2.3.1 Overview

The term ***neural network*** (or, more precisely, ***artificial neural network***) refers to a class of computational models underpinned by elements called ***nodes*** or ***neurons***. These elements are inspired by neurons in the brain. Early neurons performed addition and subtraction operations on binary signals, then thresholded the result with a binary step function [23]. In a more modern context, a neuron performs some linear or affine transformation on a set of n inputs $\mathbf{x}$ to produce a scalar, then thresholds the result with an arbitrary activation function f to produce another scalar $\hat{y}$. This is represented in Equation 2.1.

$$\hat{y} = \text{neuron}(\mathbf{x}) = f(\ell(\mathbf{x})), \ell \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}) \tag{2.1}$$

Artificial neural networks comprise sets of neurons whose inputs and outputs are connected to one another. In a relatively simple kind of neural network known as a ***feedforward neural network***, this is realized by organizing sets of neurons into layers and connecting the outputs from one layer to the next. Alternatively, this can be viewed as layers of outputs/inputs; this is shown in Figure 2.3, which displays a feedforward neural network with three layers of neurons. A neural network with a small number of layers is known as a ***shallow neural network***.

On this note, without the activation functions, the aggregate effect of combining these transformations is also a linear or affine transformation. As such, activation functions almost universally tend to be nonlinear to allow for more varied functionality (and non-affine– we will consider relevant statements about linearity to also reflect affineness from this point on).

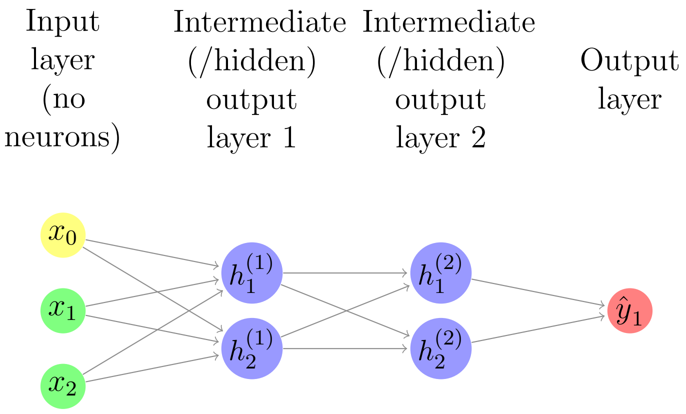


Figure 2.3: A feedforward neural network representing input/output layers. Each layer of neurons connects the inputs/outputs. Every layer past the input layer has an associated set of neurons that produces its contents. The model has three layers of neurons.

Each artificial neural network neuron's linear transformation is represented as a collection of parameters $\mathbf{w}_{\text{in\_size}}$, where `in_size` is the number of signals/edges feeding into the node. In a ***fully connected*** network, such as the feedforward network depicted in Figure 2.3, all nodes from one layer feed into the nodes of the next layer, meaning that the total number of weights feeding into a given layer of size m from the prior layer of size n is $m \cdot n$. Representing each layer as a vector then leads to the natural representation of these weights as an $m \times n$ matrix, where matrix multiplication on the prior layer leads to the current one. The representation of this process for the neural network in Figure 2.3 is demonstrated in Equation 2.2. Correspondingly, feedforward neural networks are represented by a series of `num_layers` matrices.

$$\begin{aligned}
\mathbf{h}^{(1)} &= f\left(\mathbf{M}_1\mathbf{x}\right), \mathbf{M}_1 \in \mathcal{M}_{2\times 3}(\mathbb{R}) \\
\mathbf{h}^{(2)} &= f\left(\mathbf{M}_2\mathbf{h}^{(\mathbf{1})}\right), \mathbf{M}_2 \in \mathcal{M}_{2\times 2}(\mathbb{R}) \\
\hat{y} &= f\left(\mathbf{M}_3\mathbf{h}^{(\mathbf{2})}\right), \mathbf{M}_3 \in \mathcal{M}_{1\times 2}(\mathbb{R})
\end{aligned} \tag{2.2}$$

Another important property of activation functions is that they be continuously differentiable. In the event that they are, all operations involved in the computation of the network are differentiable. By using continuously differentiable functions on the output of the neural network to produce a measure of fitness L for whatever you want the neural network to achieve, you can construct a gradient for L with respect to the weights of the matrix and update them to maximize or minimize $L(\text{Neural Network}(\mathbf{x}))$ (or, more technically in the case of certain loss functions, $L(\mathbf{x})$). This procedure is known as ***backpropagation***, and it allows neural networks to be tuned for a wide variety of tasks. We say the model is ***trained on a set of samples*** $\mathcal{X} = \{\mathbf{x}\}$ when it learns weights from those samples via backpropagation.

2.3.2 Embedding

One can use these models to construct vector representations of arbitrary discrete objects or sets of objects in the real world—types of tea, crackers, words, or anything you can derive a loss function for. The simplest way of doing so for a set of n objects is to construct a model that has a single neuronal layer of $d \cdot n$ neurons, where d is the desired number of dimensions for each object. This imposes an underlying matrix $\mathbf{M} \in \mathcal{M}(n, d)$. Each row of $\mathbf{M}$ corresponds to one of the objects, and at training time an object's vector is retrieved as in Equation 2.3.

$$v_{\text{object}_2} = \underbrace{\begin{array}{c} \text{object}_1 \\ \text{object}_2 \\ \vdots \\ \text{object}_n \end{array} \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1,d} \\ a_{21} & a_{22} & \cdots & a_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,d} \end{bmatrix}}_{\text{Object embedding } \mathbf{M}} \cdot \underbrace{\begin{array}{c} \text{object}_1 \\ \text{object}_2 \\ \vdots \\ \text{object}_n \end{array} \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}}_{\text{One-hot encoding vector}} \tag{2.3}$$

A loss function can then be set up on these vectors for some goal, and the model will learn vectors with semantic information connecting the objects to that goal. Models with sufficiently small vectors may encode information in a manner that is human-understandable, resulting in an ***interpretable model***, or models whose operations are understandable in some way.

2.4 Hebart et al. Model

We extensively use the prior work of Hebart et al. [13], who constructed a model with a shallow neural network with one layer of neurons that have an identity activation function ($f(x) = x$). This model was trained to perform an object-similarity judgment task and obtain a vector embedding for each of 1854 objects. The neural network is interpretable, with vectors for each object and each dimension of each vector corresponding to a human-understandable type of similarity. It was created as part of the THINGS initiative, an initiative to tie research projects' results through a common set of imageable objects [15].

2.4.1 THINGS Object Dataset

The underpinning of the THINGS initiative is the THINGS object dataset [14], a collection of 1854 objects with associated metadata, including English names, WordNet [18] senses, and representative images.

2.4.2 Gathered Odd-One-Out Data

Hebart et al. constructed questions for a task known as the odd-one-out task, wherein a respondent is presented with images of three objects and asked to identify the least similar among them. The objects and associated representative object images of the THINGS dataset were used for this.

They collected these responses from adults of various ages via Amazon's crowd-sourced response service Mechanical Turk. The resulting dataset consists of 4.7 million (4 699 160) adult samples.

2.4.3 Model Architecture

The neural network's architecture is that of a single-neuronal-layer embedding network that uses the identity activation function; see Section 2.3.2. It was initially initialized with an 1854×90 embedding matrix, but by using a loss function that encourages matrix sparseness (see Section 2.4.4), after training and removing columns of zeros, it is an 1854×49 embedding matrix. The model's rows correspond with the 1854 objects of the THINGS dataset, while the columns correspond with human-identifiable object-similarity feature labels, such as "long/thing" and "animal-related". An example vector retrieval for the THINGS object "abacus" that also illustrates the layout of the model embedding is given in Equation 2.4.

$$v_{\text{abacus}} = \underbrace{\begin{array}{cc} & \begin{array}{cccc} \text{metallic} & \text{food-related} & \dots & \text{cylindrical} \end{array} \\ \begin{array}{c} \text{aardvark} \\ \text{abacus} \\ \\ \text{zucchini} \end{array} & \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1,49} \\ a_{21} & a_{22} & \cdots & a_{2,49} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1854,1} & a_{1854,2} & \cdots & a_{1854,49} \end{bmatrix} \end{array}}_{\text{Object-similarity embedding}} \cdot \underbrace{\begin{array}{cc} \begin{array}{c} \text{aardvark} \\ \text{abacus} \\ \vdots \\ \text{zucchini} \end{array} & \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} \end{array}}_{\text{One-hot encoding vector}} \tag{2.4}$$

In order to compare the objects "pencil" and "baton," the model calculates their similarity by taking the dot product of the associated vectors yielded by the model as in Equation 2.5, summing up the products of each of the dimensions.

$$\begin{aligned} \text{sim}(\text{pencil}, \text{baton}) &= v_{\text{pencil}} \cdot v_{\text{baton}} \qquad (2.5) \\ &= v_{\text{pencil}}[1] \cdot v_{\text{baton}}[1] + \cdots \\ &\quad + v_{\text{pencil}}[\texttt{long/thin dim.}] \cdot v_{\text{baton}}[\texttt{long/thin dim.}] \\ &\quad + v_{\text{pencil}}[\texttt{animal-related dim.}] \cdot v_{\text{baton}}[\texttt{animal-related dim.}] \\ &\quad + \cdots + v_{\text{pencil}}[49] \cdot v_{\text{baton}}[49] \end{aligned}$$

Consider, without loss of generality, the three objects "abacus," "calculator," and "dog." The odd one out among these three objects is computed by first taking an unnormalized score for each as being the calculated similarity between the other two, as in Equation 2.6, then taking the softmax of these similarities to produce probabilities, as in Equation 2.7. Finally, the predicted odd one out is the object with the highest associated probability, as in Equation 2.8.

$$\begin{gathered} z_{\text{abacus}} = \text{sim}(\text{calculator}, \text{dog}) = v_{\text{calculator}} \cdot v_{\text{dog}} \\ \text{given the comparison set } \{\text{abacus}, \text{calculator}, \text{dog}\} \end{gathered} \tag{2.6}$$

$$\text{P}(\text{abacus is the odd-one-out}) = \sigma\left(\mathbf{z}\right)_{\text{abacus}} = \frac{e^{z_{\text{abacus}}}}{e^{z_{\text{calculator}}} + e^{z_{\text{dog}}} + e^{z_{\text{abacus}}}} \tag{2.7}$$

$$\text{prediction}(\{\text{abacus}, \text{calculator}, \text{dog}\}) = \underset{i \in \{\text{abacus},\text{calculator},\text{dog}\}}{\arg\max} \quad \sigma\left(\mathbf{z}\right)_i \tag{2.8}$$

2.4.4 Training and Loss

The model was trained using cross-entropy loss with a sum-of-vector-ℓ^1-norm penalty on the neural network embedding to encourage sparsity. The equation for cross-entropy loss is given below:

$$\begin{aligned} H(q,p)_{\substack{\text{object set is } \{i,j,k\}, \\ \text{k is the odd-one-out}}} &= \sum_{c \in \{i,j,k\}} q_{c \text{ is the odd-one-out}} \cdot \ln(p_{c \text{ is the odd-one-out}}) \\ &= -\ln\left(p(c_{\text{odd-one-out}})\right) \\ &= -\ln\left(\sigma\left(\mathbf{z}\right)_c\right) = -\ln \frac{e^{z_k}}{e^{z_k} + e^{z_j} + e^{z_i}} \\ &= -\ln \frac{e^{\vec{x}_i \vec{x}_j}}{e^{\vec{x}_i \vec{x}_j} + e^{\vec{x}_i \vec{x}_k} + e^{\vec{x}_j \vec{x}_k}} \end{aligned} \tag{2.9}$$

where

- H is the cross-entropy loss function
- i, j, k denote the three objects of a triplet, where k is the true odd-one-out

- z_c where $c \in \{i, j, k\}$ represents the similarity between the pair $\{i, j, k\} \setminus \{c\}$ (see Equation 2.6)
- $\mathbf{z} = \{z_i, z_j, z_k\}$
- σ is the softmax function (see Equation 2.7)
- q is the probability of an object being the odd one out (so 100% for the human-labelled odd-one-out, 0% for any other object)
- p is the estimated probability the model gives that a given object is the odd-one-out
- x_c is the embedding vector for object c

In this manner, the full loss expression for a set of samples $\mathcal{X}$ and embedding matrix $\mathbf{M}$, where $\mathcal{X}$ is organized such that the third object is always the odd-one-out, is given by Equation 2.10.

$$-\sum_{\{i,j,k\}\in\mathcal{X}} \ln \frac{e^{\vec{x}_i\vec{x}_j}}{e^{\vec{x}_i\vec{x}_j} + e^{\vec{x}_i\vec{x}_k} + e^{\vec{x}_j\vec{x}_k}} + \sum_{\mathbf{v}_i\in\mathbf{M}} \ell^1(\mathbf{v}_i) \tag{2.10}$$

Chapter 3

Adult Age and Dimensional Similarity Preferences

We examine and compare types of similarity preferences determined by the dimensions of an interpretable embedding, referred to as dimensional likeness preferences, exhibited by 25–35- and 50–60-year-olds when performing an object-comparison task.

3.1 Overview

As people age, they adjust how they compare objects [9]. We employ an interpretable, machine-learned model to examine differences in the preferences for these associations displayed by two age groups: 25–35- and 50–60-year-olds. Given a model with interpretable object-likeness features trained to perform object comparison, we modify the model to allow for reweighting those features. We then search for optimal feature-reweighting parameters for this modified model for the two age groups. The new reweighting parameters are taken as indications of the groups' preferences for the model features. Finally, we display these quantified preferences and use them to determine differences between each age group in their object-comparison likeness preferences.

3.2 Question Examined

How does age impact object-likeness preferences in comparing objects?

Question Do adults aged 25–35 and 50–60 prioritize object-comparison likenesses differently when comparing objects?

Procedure We first obtain measures of object-likeness prioritization for 25–35-year-olds and 50–60-year-olds. We do this by modifying an object-comparison-performing model whose dimensions correspond with object-comparison likenesses and have human-identifiable labels. Our modified model has a layer of "preference weights" that rescale the prominence of the likenesses in the original model when determining object similarities. We learn these preference weights for each age group. Finally, we perform statistical testing to see which preference weights significantly differ between groups.

3.3 Introduction

Do younger adults (25–35) and older adults (50–60) have different priorities when determining two objects' similarity?

Comparing objects is a multifaceted task. Correspondingly, when people judge the degree of similarity of two objects, they consider multiple avenues of comparison. Consider a fishing pole and a fishing spear, for example. One might judge them to be similar based on their utility in catching animals for food, but one might also judge them to be similar based on their shape. Ways of determining object similarity are known as ***semantic relations***. Individuals can vary in their perception of two objects' similarity, but broad trends emerge when looking across different objects and at groups of people.

There are many kinds of semantic relations. Here, we introduce two broad kinds of semantic relations in the literature for motivation: taxonomic and thematic relations. Taxonomic relations are feature-focused, such as the "four-legged" likeness between dogs and bears. Meanwhile, thematic relations describe those that are context-focused, such as the "pet-related" likeness between dogs and bones. Prior work on taxonomic and thematic relations has found age-related differences in people's prioritization of each when determining object similarity. Namely, older adults (66–85) exhibit a relatively stronger preference for thematic (context-based) relations over taxonomic (feature-based) relations as compared to younger adults (17–23) [9]. Considering the broadness of these two classes of relations, we might also expect other, more specific kinds of similarity preferences, which we refer to as ***likenesses*** to avoid confusion with other terms in the literature, to change with age as well.

What kinds of likenesses should we examine? Individual likenesses themselves can be manifold, and many likenesses are partitionable into more specific ones. For example, "lake-related" can be narrowed into "saltwater lake-related" or "freshwater lake-related." These specific likenesses can also be combined to build broader likenesses: "saltwater lake-related" and "ocean-related" compose "saltwater body of water-related." Note that despite "lake-related" and "saltwater body of water-related" overlapping (as demonstrated by "saltwater lake-related"), they also relate distinct objects (any two freshwater lakes and any two oceans, for example). With a plethora of likenesses to choose from and significant overlap between many, selecting which likenesses to examine and determining to what extent arbitrary objects are related under each is complicated.

Fortunately, recent advances in modelling object-similarity preferences give us both a way of choosing likenesses and of quantifying them. The authors Hebart et al. collected millions of responses from people for an odd-one-out object-comparison task (wherein images of three objects are presented and participants are asked to identify the "odd one out") (see Section 2.4). With the resulting large-scale dataset, they then trained a sparse, single-layer computational model to perform the odd-one-out task on the objects used. This method of training leads to each object having scores along a small number of dimensions. Importantly, these dimensions end up being interpretable, with each corresponding with a type of human-identifiable likeness (such as “long/thin” or “body part-related”). Thus, the model gives us a set of specific types of similarity, or ***dimensional likenesses***. It also gives us numeric scores for objects vis-à-vis those likenesses encoded in the model’s dimensions; we refer to those dimensions as ***likeness dimensions***.

In this chapter, we use these dimensional likenesses to analyze the priorities of 25–35- and 50–60-year-olds in judging similarity when performing the odd-one-out task. We identify these priorities relative to the broader population, then compare these priorities between age groups. We accomplish this by modifying the architecture of the odd-one-out-predicting model to contain a layer comprising a vector $\mathbf{w}$ of ***likeness-preference weights***. Each element $\mathbf{w}_i$ of the vector corresponds with one of the dimensional likenesses used in odd-one-out prediction. These weights rescale the vector embeddings learned from some original training dataset $\mathbf{S_1}$ to optimize performance for some new dataset $\mathbf{S_2}$. Each weight $\mathbf{w}_i$ then indicates the importance that the corresponding dimensional likeness has for $\mathbf{S_2}$ relative to $\mathbf{S}_1$. We approximate these likeness-preference weights for each age group. Finally, we perform statistical testing on the differences in these likeness-preference weights across age groups to determine age-related discrepancies in object comparison.

3.4 Methodology

3.4.1 Overview

This overview gives a mid-level look at our methodology, delineated explicitly in the following sections starting with Section 3.4.2. In this chapter, we focus on identifying connections between adult age and preference for object likenesses. In particular, we concentrate on two age groups, 25–35- and 50–60-year-olds, and a prior set of 49 quantified object likenesses. To determine differences between these age groups along these likenesses, we quantify preferences for each age group regarding each of these likenesses. We then compare those preferences to determine statistically significant differences. Below, we describe the decisions underlying this in several parts:

1. Prior data
2. Choice of age groups
3. Choice of object likenesses and prior model
4. Preference-learning modified model architecture
5. Determining optimal preferences
6. Statistical testing

First, we discuss a prior object-comparison response dataset. The authors Hebart et al. collected millions of responses to the odd-one-out task, where respondents are presented with three object images and asked to identify the least similar amongst them [13]. The object images in question are from the THINGS object dataset, a collection of 1854 objects with associated metadata [14]. We shall henceforth refer to these collected responses as the THINGS odd-one-out dataset. Additional information on this dataset is located in Section 2.4.1.

Second, we discuss our choice of age groups. The responses of the THINGS odd-one-out dataset have respondent-age annotations. The distribution of the responses is multimodal, with peaks at ages 30 and 56. We choose 25–35- and 50–60-year-olds in order to take advantage of as many triplets as possible while maintaining two distinct age groups. This distribution can be found in Section 3.4.3

Third, we discuss the prior model and the choice of object likenesses to examine when comparing age groups. From Hebart et al., we have the model with 49 likeness-encoding features—i.e., ***likeness dimensions***—that quantify a set of abstract ***di-***

mensional likenesses with human labels. This model performs the odd-one-out task after training on the THINGs odd-one-out dataset. The model's architecture is a shallow neural network. It embeds each of the THINGs objects into vector form, then performs a dot-product operation between two of them to judge their similarity. The authors train the model sparsely, and as mentioned each resulting embedding dimension corresponds with a human-identifiable likeness, such as "long/thin" or "body part-related." For our purposes, this gives us 49 dimensional likenesses with which to work. The architecture of this model is explained in Section 3.4.5.

Fourth, we discuss how we learn a group's preferences for these dimensional likenesses. To do this, we modify Hebart et al.'s model to have a ***likeness-preference layer***. This layer contains a ***likeness-preference weight*** for each likeness dimension that rescales that dimension's values when the model performs the odd-one-out task. The model may be initially trained while ignoring this layer and fixing it to a set of ones, although we use Hebart et al.'s existing embedding rather than doing so. Then, we fix the embedding and allow the likeness-preference layer to vary[1]. Training these likeness-preference weights on a set of responses gives us a measure of how much more effectively that relational dimension contributes to accurately modelling the odd-one-out choices in those responses. We interpret these weights as a proxy for the respondents' object-likeness preferences. A more explicit description, as well as some intuition, can be found in Section 3.4.6

Fifth, by training each likeness-preference weight of the likeness-preference layer on a particular set of odd-one-out responses, we then learn preferences for each dimensional likeness specific to those responses. Using responses from a given age group, we can then observe that age group's preferences for each dimensional likeness relative to the overall population. Feeding the responses of each age group to the modified model, we

[1] While we do not elaborate on this here, we technically learn the weight for one dimension at a time rather than all 49 likeness-preference weights at once due to the embedding matrix not being linearly independent. See Section 3.4.7 for specific documentation of this.

learn measures of their respective preferences for each of the 49 dimensional likenesses. Further details on obtaining these weights are located in Section 3.4.7

Sixthly and lastly, we discuss our testable experiment. We load Hebart et al.'s model embedding into our modified model architecture that additionally incorporates likeness-preference weights. Without loss of generality, we select a dimension for which to learn an optimal likeness-preference weight. We learn the optimal preferences for each age group on the full sets of each group's response data and take the difference. Finally, we combine the two groups of responses and learn random preference differences, using that to perform a bootstrapped statistical test on whether the difference between the age groups' preferences exceeds chance. A formal delineation of these tests is given in Section 3.4.8

Below, we go into further detail about each of these steps.

3.4.2 Prior Data and Cleaning

To start, we have a large number of existing responses to an object-comparison task courtesy of Hebart et al. [13] They collected these responses for the object-comparison task known as the "odd-one-out" task, where a respondent is presented with images of three objects and asked to identify the least similar among them. They collected these responses from adults of various ages via Amazon's crowdsourced response service Mechanical Turk.

The resulting dataset consists of 4.7 million (4 699 160) adult samples, of which 3.26 million (3 259 599) have age annotations. After filtering out the 7340 samples that have a user-entered age of over 110 and the 240 that have an age under 18, we are left with 3.25 million (3 252 020) adult responses to work with.

3.4.3 Choice of Age Groups

In order to determine differences in object-similarity preferences by age, we need results from different age groups to compare. The age-annotated responses from the object-comparison dataset from participants of ages 18 and up form the multimodal distribution displayed in Figure 3.1.

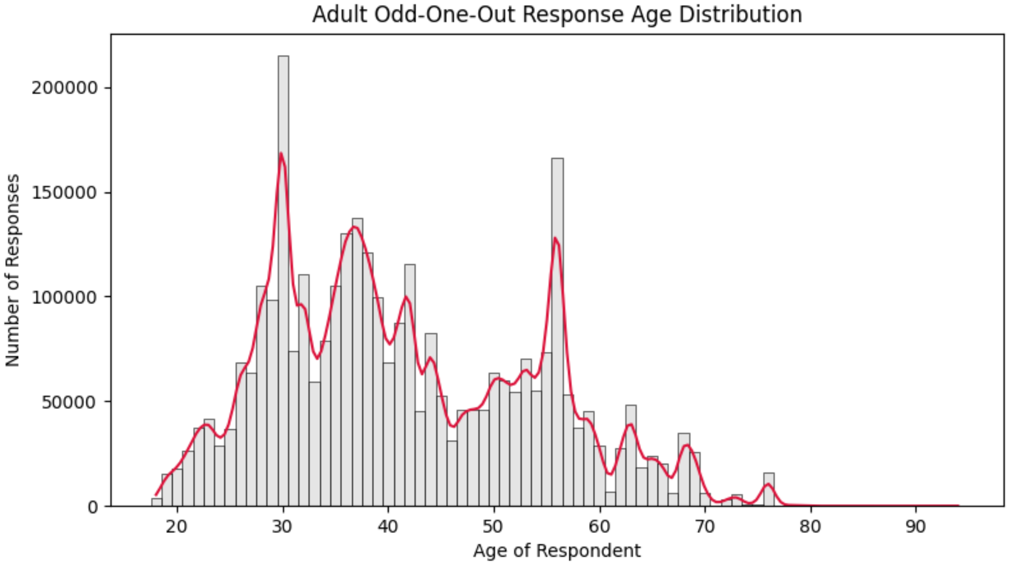


Figure 3.1: The distribution of response ages in the THINGS response dataset.

This distribution has peaks at ages 30, 37, and 56. Because we want age groups sufficiently far apart, we ignore the peak at 37 and center our age groups around the remaining peaks by taking the age ranges of 25–35 and 50–60 as our respondent groups. For the age group 25–35, this gives us 1.02 million (1 015 960) responses, while for the age group 50–60, we have 710 thousand (708 420) responses.

3.4.4 Choice of Object Likenesses

Concerning the choice of types of object similarity, or ***likenesses***, on which to compare the preferences between our two age groups, we use the likenesses encoded in an existing model—namely, the model that Hebart et al. developed to perform the odd-one-out object-comparison task [13]. The model contains a vector of 49 object-likeness scores for each of the 1854 objects of the THINGS dataset [14]. Each score corresponds with a human-identifiable object likeness, such as “long/thin” or “body part-related,” and indicates the degree to which the object is similar to others with regard to that likeness. We henceforth refer to the object likenesses that the dimensions encode as ***dimensional likenesses*** and the dimensions themselves as ***likeness dimensions***.

The likeness dimensions are encoded in the model as an 1854 × 49 matrix. Each column corresponds with a human-labelled summary of the dimensional likeness, and each row denotes a specific object’s scores along the associated likeness dimensions. This is illustrated in Equation 3.1.

$$\underbrace{\begin{array}{rc} & \begin{array}{cccc} \text{metallic} & \text{food-related} & \dots & \text{cylindrical} \end{array} \\ \begin{array}{r} \text{aardvark} \\ \text{abacus} \\ \\ \text{zucchini} \end{array} & \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,49} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,49} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1854,1} & a_{1854,2} & \cdots & a_{1854,49} \end{bmatrix} \end{array}}_{\text{Learned object-similarity embeddings}} \tag{3.1}$$

As previously mentioned, these dimensional likenesses have human-identifiable labels, such as “long/thin” or “body part-related.” When the model compares two objects, it takes the dot product of their constituent vectors. More descriptively, it multiplies their respective scores along each dimension and sums the products to form a single object-similarity score, as illustrated in Equation 3.2.

$$\begin{aligned}\text{sim}(\text{pencil}, \text{baton}) &= v_{\text{pencil}} \cdot v_{\text{baton}} \\ &= v_{\text{pencil}}[1] \cdot v_{\text{baton}}[1] + \cdots \\ &\quad + v_{\text{pencil}}[\texttt{long/thin dim.}] \cdot v_{\text{baton}}[\texttt{long/thin dim.}] \\ &\quad + v_{\text{pencil}}[\texttt{animal-related dim.}] \cdot v_{\text{baton}}[\texttt{animal-related dim.}] \\ &\quad + \cdots + v_{\text{pencil}}[49] \cdot v_{\text{baton}}[49]\end{aligned} \tag{3.2}$$

As such, the product of two objects' scores for a dimension gives another score, one that indicates the extent to which the corresponding likeness contributes to the two objects' overall similarity.

3.4.5 Original Model

The original Hebart et al. model, which we will be modifying for our purposes, performs the odd-one-out-among-three object-comparison task through three main steps.

First, it converts each object into a vector representation by means of its learned embedding matrix. The details of the training are found in Section 2.4.4. A sample vector embedding for "abacus" is given in expression Equation 3.3.

$$v_{\text{abacus}} = \underbrace{\begin{array}{cc} & \begin{array}{cccc} \text{metallic} & \text{food-related} & \dots & \text{cylindrical} \end{array} \\ \begin{array}{c} \text{aardvark} \\ \text{abacus} \\ \\ \text{zucchini} \end{array} & \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1,49} \\ a_{21} & a_{22} & \cdots & a_{2,49} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1854,1} & a_{1854,2} & \cdots & a_{1854,49} \end{bmatrix} \end{array}}_{\text{Fixed object-similarity embeddings}} \cdot \underbrace{\begin{array}{cc} \begin{array}{c} \text{aardvark} \\ \text{abacus} \\ \vdots \\ \text{zucchini} \end{array} & \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} \end{array}}_{\text{One-hot encoding vector}} \tag{3.3}$$

Second, the model computes a scalar similarity between each pair of objects by taking the dot product of the objects' embedding vectors. This scalar similarity depends solely on the other two objects being considered. It can be interpreted as an unnormalized score correlated with the chance of the out-of-pair object being the odd-one-out. This is expressed in Equation 3.4.

$$\begin{gathered} z_{\text{abacus}} = \text{sim}(\text{calculator}, \text{dog}) = v_{\text{calculator}} \cdot v_{\text{dog}} \\ \text{given the comparison set } \{\text{abacus}, \text{calculator}, \text{dog}\} \end{gathered} \tag{3.4}$$

Finally, the model transforms the raw scores into probabilities by taking the softmax of the raw scores. As a reminder, the softmax function transforms any number of real-valued scalars into a probability distribution. This is illustrated in Equation 3.5. The model takes the object granting the highest among these probabilities to be its predicted odd-one-out, as demonstrated in Equation 3.6.

$$\text{P}(\text{abacus is the odd-one-out}) = \sigma(\mathbf{z})_{\text{abacus}} = \frac{e^{z_{\text{abacus}}}}{e^{z_{\text{calculator}}} + e^{z_{\text{dog}}} + e^{z_{\text{abacus}}}} \tag{3.5}$$

$$\text{prediction}(\{\text{abacus}, \text{calculator}, \text{dog}\}) = \underset{i \in \{\text{abacus},\text{calculator},\text{dog}\}}{\arg\max} \sigma(\mathbf{z})_i \tag{3.6}$$

3.4.6 Modified Model

In order to compare object-similarity preferences between age groups, we need measures of those preferences. We accomplish this by modifying the original model architecture to incorporate what we call ***likeness-preference weights***. These create an additional layer of the model that transforms any retrieved embedding vector via element-wise rescaling.

Computationally, we accomplish this by introducing a 49-dimensional vector into the model, where each entry is a ***likeness-preference weight***. This vector is embedded into the diagonal of a 49 × 49 matrix, which rescales any object-similarity vector retrieved from the embedding matrix. In this manner, each of the 49 likeness-preference weights multiplies the associated likeness dimension of a given original embedding vector. This is illustrated in Equation 3.7.

$$\begin{gathered}
v'_{\text{abacus}} = \underbrace{\begin{array}{cccc} \text{aardvark} & \text{abacus} & \dots & \text{zucchini} \\ \left[\quad 0\right. & 1 & \cdots & \left.0 \quad\right] \end{array}}_{\text{One-hot object-choosing vector}} \cdot \\
\underbrace{\begin{array}{cccc} \text{metallic} & \text{food-related} & \dots & \text{cylindrical} \\ \hline a_{1,1} & a_{1,2} & \cdots & a_{1,49} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,49} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1854,1} & a_{1854,2} & \cdots & a_{1854,49} \end{array}}_{\text{Fixed object-similarity embeddings}} \cdot \underbrace{\begin{array}{cccc} \text{metallic} & \text{food-related} & \dots & \text{cylindrical} \\ \hline c_1 & 0 & \cdots & 0 \\ 0 & c_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & c_{49} \end{array}}_{\substack{\text{Diagonally embedded} \\ \text{reweighting vector} \\ \text{of likeness-preference weights}}}
\end{gathered} \tag{3.7}$$

The model then performs odd-one-out predictions by taking the object with the maximum associated softmax probability as in Section 3.4.5. Due to the dual facts that one: the likeness-preference weights transform the object-embedding vectors, and two: the method of prediction is the same, our model is equivalent to the original, unmodified Hebart et al. model when the 49 likeness-preference weights are set to one. While the model embedding was trained on some set of odd-one-out responses S_1, the preference-reweighting layer is trained on some other set of odd-one-out responses S_2. We give intuition regarding this in Section 3.4.7.1.

3.4.7 Likeness-Preference Weight Optimization

3.4.7.1 Likeness-Preference Weight Intuition

Our modified model uses the object embedding of Hebart et al. trained on a set of odd-one-out responses S_1. The likeness-preference weights, meanwhile, are designed to be trained on a separate set of responses. Were the likeness-preference weights to be trained on S_1, the expected optimal values for them would be those producing the original model; more specifically, they would all have a value of one. Any deviations from this would be for one of two reasons: marginal effects resulting from suboptimality of training the original embedding, or likenesses between embedding dimensions resulting in multiple optima in the original embedding.

As such, we instead train the likeness-preference weights on some alternative set of responses S_2. As likeness-preference weights reweight the columns of the original model embedding trained on S_1, any likeness-preference weights learned are relative to the choice of those columns. For both scenarios below, assume uncorrelated columns:

- Consider S_1 and S_2 from populations with identical object-comparison preferences that are asked about the same objects. The expected likeness-preference weights are within some small marginal difference from one.

- Consider S_1 and S_2 from populations with different object-comparison preferences that were asked to perform the odd-one-out task on the same sets of three objects. The expected likeness-preference weights for S_2 are beyond one in accordance with the magnitude of those differences in preferences. In other words, these preferences are relative to those of S_1.

Note that the columns of the Hebart et al. object embedding are correlated. We can bypass this issue by training the likeness-preference weights independently, as described below.

3.4.7.2 Isolating Likeness-Preference Weights

Rather than optimizing all likeness-preference weights for the model simultaneously, we optimize each likeness-preference weight individually, leaving the others fixed. We isolate the process of training each likeness-preference training because the likeness dimensions of the model share information, and so an increase in one likeness-preference weight can result in a decrease in another. While this limits the overall performance gain that we can achieve over the original mode, this is inconsequential: we are primarily interested in learning optimal likeness-preference weights of each dimension in isolation rather than optimizing their collective utility in model performance.

3.4.7.3 Loss Function

We learn likeness-preference weights by minimizing cross-entropy loss. Specifically, the cross-entropy $H(q,p)$ of the model prediction probability p relative to the theoretical actual prediction probability q is given by

$$\begin{aligned} H(q,p)_{\substack{\text{object set is } \{i,j,k\},\\ \text{k is the odd-one-out}}} &= \sum_{c\in\{i,j,k\}} q_{c \text{ is the odd-one-out}} \cdot \ln(p_{c \text{ is the odd-one-out}}) \\ &= -\ln\left(p(c_{\text{odd-one-out}})\right) \\ &= -\ln\left(\sigma\left(\mathbf{z}\right)_c\right) = -\ln\frac{e^{z_k}}{e^{z_k}+e^{z_j}+e^{z_i}} \\ &= -\ln\frac{e^{\vec{x}_i\vec{x}_j}}{e^{\vec{x}_i\vec{x}_j}+e^{\vec{x}_i\vec{x}_k}+e^{\vec{x}_j\vec{x}_k}} \end{aligned} \tag{3.8}$$

where

- H is the cross-entropy loss function
- i, j, k denote the three objects of a triplet, where k is the true odd-one-out
- z_c where $c \in \{i,j,k\}$ represents the similarity between the pair $\{i,j,k\} \setminus \{c\}$ (see Equation 3.5)
- $\mathbf{z} = \{z_i, z_j, z_k\}$
- σ is the softmax function (see Equation 3.5)
- q is the probability of an object being the odd one out (so 100% for the human-labelled odd-one-out, 0% for any other object)
- p is the estimated probability the model gives that a given object is the odd-one-out

- x_c is the reweighted embedding vector for object c; i.e. $\left[x_{c0}^{(1)}, x_{c0}^{(2)}, ..., w_d x_{c0}^{(d)}, ..., x_{c0}^{(49)}\right]$, where x_{c0} is the unreweighted embedding vector for object c and d is the unique dimension being reweighted (as per Section 3.4.7.2)

The gradient of this with respect to the reweighting dimension value of the dimension being considered can be found in Section 3.6.

Across a set of responses $\mathcal{S}$, we take the average cross-entropy loss, $\frac{1}{|\mathcal{S}|} \sum_{s \in \mathcal{S}} H(q, p)$.

3.4.7.4 Convexity and Stochasticity

The cross-entropy loss (see Section 3.4.7.3) is highly stochastic, depending heavily on the batch of samples used. This stochasticity holds even up to a vast number of samples (on the order of one million). Nevertheless, it is consistently convex on a per-batch level. Samples of the loss curve when varying the likeness-preference weight of the first dimension of the model across five different fixed batches are given in Figure 3.2.

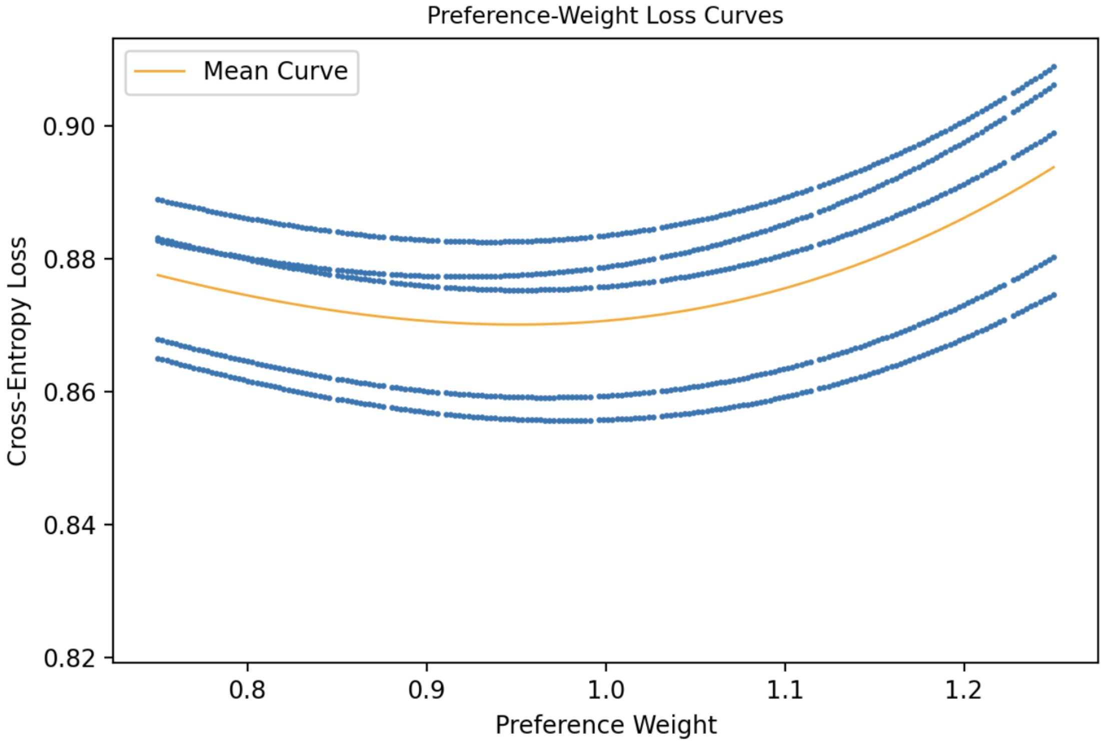


Figure 3.2: Five sample loss curves achieved by varying the likeness-preference weight of Dimension 1 under a fixed batch of responses for each. The orange line represents the mean of the curves. The x-axis gives the value of the likeness-preference weight within the modified model, while the y-axis denotes the resulting loss of the modified model on a fixed batch of responses.

Meanwhile, due to the stochasticity of the model's loss when the batch is varied, the resulting loss curve is not convex, and naïve cross-batch optimization methods result in subpar results. We give an example loss curve calculated across different batches in Figure 3.3.

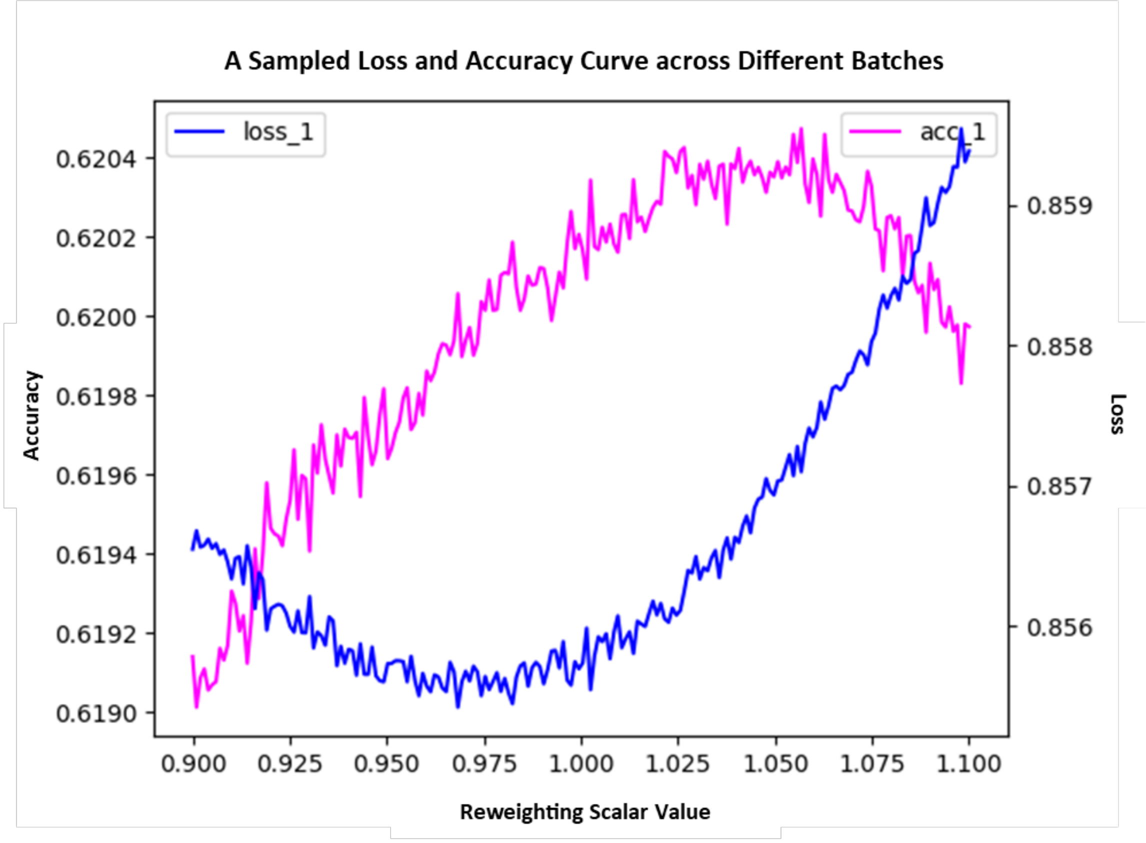


Figure 3.3: A noisy loss curve sampled across many batches for Dimension 1. The x-axis is the value of the likeness-preference weight for Dimension 1 within the modified model, while the y-axis is the loss incurred by the model. The blue line is the curve taken by varying the likeness-preference weight and recording the model loss on random batches. Notice the non-convexity of the loss curve, which stands in contrast to the convex loss curves of Figure 3.2, where each curve was obtained on a single batch.

3.4.7.5 Choice of Likeness-Preference-Weight Optimization Method

As described in the previous section, due to the stochastic nature of the loss of the model with respect to different batches, naïve cross-batch learning methods have convergence issues. However, we can still optimize the likeness-preference weight on individual, fixed batches. As we 1) have the derivative 2) know the loss curve for individual batches is continuous and convex, we use Alefield et al.'s Algorithm 748 [24] and find the zero of the derivative to identify the minimum loss for a batch of responses.

3.4.8 Testing and Sampling

3.4.8.1 Bootstrapped Statistical Test

Recall that we wish to learn preferences for each age group. We use the test procedure outlined in [25, Section 4.4.1] for determining whether the difference in a statistic between two age groups is significant. More specifically, the procedure is as follows for a given dimension.

Consider a statistic f on a given sample of responses $\mathbf{s}$, as well as large samples $\mathbf{S}_1$ and $\mathbf{S}_2$. For our experiment, $f(\mathbf{s})$ is the optimal preference weight learned on $\mathbf{s}$, while $\mathbf{S}_1$ and $\mathbf{S}_2$ are the full sets of odd-one-out responses for 25–35-year-olds and 50–60-year-olds. We will consider a new test statistic on two samples given by the difference between their preference weights, $t(\mathbf{s}_1, \mathbf{s}_2) = f(\mathbf{s}_2) - f(\mathbf{s}_1)$.

1. First, take $t_{\text{overall}} = t(\mathbf{S}_1, \mathbf{S}_2)$.

2. Next, create a dataset $\mathbf{S}_{\text{both}}$ by merging $\mathbf{S}_1$ and $\mathbf{S}_2$.

3. Create n random subsamples with replacement, or ***bootstrapped samples***, $(\mathbf{s}_1^*, ..., \mathbf{s}_n^*)$ of $\mathbf{S}_{\text{both}}$. We actually do this twice, and get $(\mathbf{x}_1^*, ..., \mathbf{x}_n^*)$ and $(\mathbf{y}_1^*, ..., \mathbf{y}_n^*)$. In our case, $\mathbf{S}_1$ and $\mathbf{S}_2$ are of different sizes, so we make sure that each bootstrapped sample has an equal number of responses from each age group. For computational reasons, we take slightly less than the number of samples from either age group set.[2]

4. For each pair of samples $(\mathbf{x}_i^*, \mathbf{y}_i^*)$, produce $t_i^* = t(\mathbf{x}_i^*, \mathbf{y}_i^*)$.

5. Finally, compute our p-value by computing $p = \frac{\sum_{i=1}^{n} t_i^* \geq t_{\text{overall}}}{n}$, where $\geq$ is overridden to return 1 or 0 rather than true or false.

This p-value is the probability of seeing our observed difference between the preference

[2] Our setup is slightly prone to Type-II errors, as the standard error of the estimates for our smaller bootstrapped samples will be larger than if they matched the sizes of $\mathbf{S}_1$ or $\mathbf{S}_2$. We are less concerned with missing results than with reporting erroneous novel results, however.

weights of $\mathbf{S}_1$ and $\mathbf{S}_2$ under the null hypothesis that it is drawn from the mixed-group distribution of taking differences of random pairs of groups formed from the union of $\mathbf{S}_1$ and $\mathbf{S}_2$.

In practice, we only bootstrap the $f(\mathbf{s}_i^*)$ values directly (taking 2000), and subsample these to get bootstrapped differences.

3.4.8.2 Bootstrapped Confidence Interval

In addition to our statistical test, we also create a confidence interval. This confidence interval is not the same as the one implied by our statistical test, although the value 0 appearing in this confidence interval is heavily correlated with low, insignificant p-values under that test. It is instead given by the distributions of subtracting a bootstrapped difference between the preference weights of two samples from the combined age groups (a t_i^*, in the parlance of the procedure in Section 3.4.8.1) from a bootstrapped difference between the separated age groups.

More formally, consider the optimal-preference-weight statistic f on a given set of odd-one-out responses $\mathbf{s}$, as well as set of all 25–35-year-old responses $\mathbf{S}_1$ and the set of all 50–60-year-old responses $\mathbf{S}_2$. Consider the difference-based test statistic t that computes the difference between two samples; i.e., $t(\mathbf{s}_1, \mathbf{s}_2) = f(\mathbf{s}_2) - f(\mathbf{s}_1)$. For a given dimension our confidence interval is taken by doing the following:

1. First, create a dataset $\mathbf{S}_{\text{both}}$ by merging $\mathbf{S}_1$ and $\mathbf{S}_2$.

2. Next, create n random subsamples with replacement, or ***bootstrapped samples***, $(\mathbf{s}_1^*, ..., \mathbf{s}_n^*)$ of $\mathbf{S}_{\text{both}}$. We actually do this twice, and get $(\mathbf{s}_1^{*\text{mixed}_1}, ..., \mathbf{s}_n^{*\text{ mixed}_1})$ and

   1 $(\mathbf{s}_1^{*\text{mixed}_2}, ..., \mathbf{s}_n^{*\text{ mixed}_2})$. In our case, $\mathbf{S}_1$ and $\mathbf{S}_2$ are of different sizes, so we make sure that each bootstrapped sample has an equal number of responses from each age group. For computational reasons, we take slightly less than the number of samples from either age group set.

3. For each pair of samples $(\mathbf{s}_i^{*\text{mixed}_1}, \mathbf{s}_i^{*\text{mixed}_2})$, produce $t_i^{*\text{mixed}} = t(\mathbf{s}_i^{*\text{mixed}_1}, \mathbf{s}_i^{*\text{mixed}_2})$.

4. Create n random bootstraps of $(\mathbf{s}_1^{*25\text{–}35}, ..., \mathbf{s}_n^{*\,25\text{–}35})$ of $\mathbf{S}_1$.

5. Create n random bootstraps of $(\mathbf{s}_1^{*50\text{–}60}, ..., \mathbf{s}_n^{*\,50\text{–}60})$ of $\mathbf{S}_2$.

6. For each pair of samples $(\mathbf{s}_i^{*25\text{–}35}, \mathbf{s}_i^{*50\text{–}60})$, produce $t_i^{*\text{age}} = t(\mathbf{s}_i^{*25\text{–}35}, \mathbf{s}_i^{*50\text{–}60})$.

7. For each pair of bootstrapped statistics $(t_i^{*\text{age}}, t_i^{*\text{mixed}})$, take the difference δ_i^{*}=$t_i^{*\text{age}}$ - $t_i^{*\text{mixed}}$.

The intuition for this is that each δ_i^* represents the extent by which a bootstrapped difference between each age group outstrips a bootstrapped mixed difference. Therefore, if a confidence interval for δ_i contains 0, it suggests one of two things: either there are inconsistencies in which age group's preference weight is larger, or the differences in the preference weights between each age group are sometimes less than what would occur by chance.

In practice, we only bootstrap the $f(\mathbf{s}_i^*)$ values directly (taking 2000), and sample differences among these to get bootstrapped differences.

3.4.8.3 Sampling

Without loss of generality, we wish to obtain the following pieces of information for a dimension d:

1. A single optimal likeness-preference weight for each age group over all of that age group's recorded responses.

2. A collection of optimal likeness-preference weights for each age group over subsets of that age group's responses.

3. A collection of the differences between the optimal likeness-preference weights of each age group.

4. A collection of optimal likeness-preference weights over subsets of the combined

set of all 25–35 and 50–60 responses.

5. A collection of the differences between the optimal likeness-preference weights.

First, we start by computing the optimal likeness-preference weight for each age group on the full set of odd-one-out responses, which numbers 1 015 960 for the 25–35-year-olds and 708 420 for the 50–60-year-olds.

Secondly, we take 4000 bootstrapped samples of each age group's responses. For computational reasons (namely, memory limits for fast computing of the preference weights) these bootstrapped samples each have 500 000 odd-one-out responses each. We compute 4000 optimal likeness-preference weight sample estimates for each age group from these.

For details about how to compute likeness-preference weights, see Section 3.4.7.

Third, we take 200 000 sample statistic differences (differences in the optimal preference weights) between each age group (50–60-year-old preference weights minus 25–35-year-old preference weights) using the 4000 bootstrapped optimal likeness-preferences.

More formal sampling details for step three are given in the sample procurement steps of Section 3.4.8.2.

Fourth, we combine the two age groups' responses into a single dataset and take two sets of 4000 bootstrapped samples from that dataset. Computing the optimal preference weights for these, this gives us two sets of 4000 mixed-group optimal likeness-preference weight sample statistics.

Fifthly and finally, we take 200 000 sample statistic differences (differences in the optimal preferences weights) between each of the two mixed-group sets of 4000 preference-weight sample statistics.

More formal sampling details for steps four and five can be found in Section 3.4.8.2

and Section 3.4.8.1.

We use the sets of sample statistics obtained in steps 1–5 above to compute the p-value significance result of Section 3.4.8.1 and the confidence interval described in Section 3.4.8.2. We do this for each of the 49 likeness dimensions and report the results.

3.5 Results and Discussion

This section contains, for each likeness dimension, the differences between the likeness-preference weights of each age group learned from their corresponding full set of odd-one-out responses. It also details the significance of these differences as determined by our bootstrapped statistical test (Section 3.4.8.1).

Table 3.1 orders these results by likeness dimension number, while Table 3.2 orders the results by differences between the age groups' likeness-preference weights.

For per-dimension graphs of the bootstrapped distributions of the age groups' preference weights, the bootstrapped differences between each age group's preference weights, and the confidence intervals (both the one associated with the statistical test of Section 3.4.8.1 and the alternative confidence interval detailed in Section 3.4.8.2), see Appendix A.

3.5.1 Differences in Likeness Preference

Dimension	Human Description	Significant (p<0.1) Age Diff.?	Age 25-35 Preference	Age 50-60 Preference	Pref. Diff. (50-60 Min. 25-35)	p-value
1	made of metal/artificial/hard	TRUE	0.972	0.996	0.024	0.000
2	food-related/eating-related/kitchen-related	TRUE	0.948	0.920	-0.028	0.000
3	animal-related/organic	TRUE	0.898	0.911	0.013	0.011
4	clothing-related/fabric/covering	FALSE	0.993	0.980	-0.012	0.103
5	furniture-related/household-related/artifact	FALSE	0.971	0.976	0.005	0.592
6	plant-related/green	FALSE	0.885	0.890	0.005	0.554
7	outdoors-related	FALSE	0.905	0.903	-0.003	0.820
8	transportation/motorized/dynamic	TRUE	0.948	0.967	0.019	0.051
9	wood-related/brownish	FALSE	0.950	0.961	0.011	0.350
10	body part-related	TRUE	0.984	0.947	-0.037	0.000
11	colorful	TRUE	0.944	1.002	0.058	0.000
12	valuable/special occasion-related	FALSE	0.952	0.945	-0.007	0.639
13	electronic/technology	TRUE	0.948	0.973	0.024	0.044
14	sport-related/recreational activity-related	TRUE	0.967	0.996	0.029	0.031
15	disc-shaped/round	TRUE	0.896	1.035	0.139	0.000
16	tool-related	TRUE	0.896	0.971	0.075	0.000
17	many small things/course pattern	TRUE	0.901	0.970	0.069	0.000
18	paper-related/thin/flat/text-related	FALSE	0.965	0.952	-0.013	0.359
19	fluid-related/drink-related	TRUE	0.948	0.922	-0.026	0.085
20	long/thin	TRUE	0.898	1.041	0.142	0.000
21	water-related/blue	FALSE	0.917	0.907	-0.009	0.544
22	powdery/fine-scale pattern	FALSE	0.858	0.874	0.016	0.505
23	red	TRUE	0.949	1.028	0.079	0.000
24	feminine (stereotypically)/decorative	FALSE	0.933	0.893	-0.040	0.109
25	bathroom-related/sanitary	FALSE	0.943	0.908	-0.035	0.178
26	black/noble	TRUE	0.967	0.902	-0.065	0.020
27	weapon/danger-related/violence	FALSE	0.906	0.897	-0.009	0.712
28	musical instrument-related/noise-related	FALSE	0.885	0.924	0.039	0.158
29	sky-related/flying-related/floating-related	FALSE	0.870	0.901	0.031	0.269
30	spherical/ellipsoid/rounded/voluminous	TRUE	0.812	0.889	0.076	0.012
31	repetitive	TRUE	0.845	0.926	0.082	0.017
32	flat/patterned	TRUE	0.776	0.871	0.096	0.005
33	white	TRUE	0.950	1.025	0.075	0.014
34	thin/flat	FALSE	0.830	0.843	0.013	0.712
35	disgusting/bugs	FALSE	0.797	0.843	0.046	0.235
36	string-related	TRUE	0.868	0.940	0.072	0.035
37	arms/legs/skin-related	FALSE	0.879	0.924	0.045	0.187
38	shiny/transparent	FALSE	0.907	0.916	0.009	0.799
39	construction-related/physical work-related	FALSE	0.831	0.861	0.030	0.477
40	fire-related/heat-related	FALSE	0.914	0.896	-0.017	0.611
41	head-related/face-related	FALSE	0.908	0.910	0.002	0.953
42	beams-related	TRUE	0.736	0.821	0.085	0.080
43	seating-related/put things on top	FALSE	0.805	0.853	0.048	0.338
44	container-related/hollow	FALSE	0.806	0.823	0.017	0.760
45	child-related/toy-related	FALSE	0.905	0.925	0.020	0.664
46	medicine-related	TRUE	0.862	0.756	-0.105	0.056
47	has grating	TRUE	0.769	0.972	0.203	0.000
48	handicraft-related	FALSE	0.603	0.583	-0.020	0.826
49	cylindrical/conical	FALSE	0.638	0.748	0.110	0.248

Table 3.1: Optimal likeness-preference differences between age groups by dimension number (sum-of-column-scores-in-embedding ordering). Significance indicates that the observed, full-response-set difference is unlikely under the bootstrapped mixed-age-group difference distribution (see Section 3.4.8.1).

Pref. Diff. Order	Dimension	Human Description	Significant (p<0.1) Age Diff.?	Pref. Diff. (50-60 Min. 25-35)	p-value
1	46	medicine-related	TRUE	-0.105	0.056
2	26	black/noble	TRUE	-0.065	0.020
3	24	feminine (stereotypically)/decorative	FALSE	-0.040	0.109
4	10	body part-related	TRUE	-0.037	0.000
5	25	bathroom-related/sanitary	FALSE	-0.035	0.178
6	2	food-related/eating-related/kitchen-related	TRUE	-0.028	0.000
7	19	fluid-related/drink-related	TRUE	-0.026	0.085
8	48	handicraft-related	FALSE	-0.020	0.826
9	40	fire-related/heat-related	FALSE	-0.017	0.611
10	18	paper-related/thin/flat/text-related	FALSE	-0.013	0.359
11	4	clothing-related/fabric/covering	FALSE	-0.012	0.103
12	21	water-related/blue	FALSE	-0.009	0.544
13	27	weapon/danger-related/violence	FALSE	-0.009	0.712
14	12	valuable/special occasion-related	FALSE	-0.007	0.639
15	7	outdoors-related	FALSE	-0.003	0.820
16	41	head-related/face-related	FALSE	0.002	0.953
17	5	furniture-related/household-related/artifact	FALSE	0.005	0.592
18	6	plant-related/green	FALSE	0.005	0.554
19	38	shiny/transparent	FALSE	0.009	0.799
20	9	wood-related/brownish	FALSE	0.011	0.350
21	3	animal-related/organic	TRUE	0.013	0.011
22	34	thin/flat	FALSE	0.013	0.712
23	22	powdery/fine-scale pattern	FALSE	0.016	0.505
24	44	container-related/hollow	FALSE	0.017	0.760
25	8	transportation/motorized/dynamic	TRUE	0.019	0.051
26	45	child-related/toy-related	FALSE	0.020	0.664
27	13	electronic/technology	TRUE	0.024	0.044
28	1	made of metal/artificial/hard	TRUE	0.024	0.000
29	14	sport-related/recreational activity-related	TRUE	0.029	0.031
30	39	construction-related/physical work-related	FALSE	0.030	0.477
31	29	sky-related/flying-related/floating-related	FALSE	0.031	0.269
32	28	musical instrument-related/noise-related	FALSE	0.039	0.158
33	37	arms/legs/skin-related	FALSE	0.045	0.187
34	35	disgusting/bugs	FALSE	0.046	0.235
35	43	seating-related/put things on top	FALSE	0.048	0.338
36	11	colorful	TRUE	0.058	0.000
37	17	many small things/course pattern	TRUE	0.069	0.000
38	36	string-related	TRUE	0.072	0.035
39	33	white	TRUE	0.075	0.014
40	16	tool-related	TRUE	0.075	0.000
41	30	spherical/ellipsoid/rounded/voluminous	TRUE	0.076	0.012
42	23	red	TRUE	0.079	0.000
43	31	repetitive	TRUE	0.082	0.017
44	42	beams-related	TRUE	0.085	0.080
45	32	flat/patterned	TRUE	0.096	0.005
46	49	cylindrical/conical	FALSE	0.110	0.248
47	15	disc-shaped/round	TRUE	0.139	0.000
48	20	long/thin	TRUE	0.142	0.000
49	47	has grating	TRUE	0.203	0.000

Table 3.2: Optimal likeness-preference differences between age groups ordered by observed difference in full-response-set preference. Significance indicates that the observed, full-response-set difference is unlikely under the bootstrapped mixed-age-group difference distribution (see Section 3.4.8.1).

3.5.2 Discussion

As a recap: we have searched for optimal preferences for each of 49 object likenesses. We have done this for two adult age groups, ages 25–35 and 50–60, in hopes of determining population-level differences in the usages of these likenesses. The results are displayed ordered by dimension in Table 3.1 and ordered by level of difference in Table 3.2. We find statistically significant differences in the preferences for each age group for 23/49 of the dimensions at $p = 0.1$.

It is important to clarify here that while the dimensions do correspond with human-identifiable labels, they are a specific understanding of them learned on the original embedding's training population. Thus, for example, while a group having a lower preference score for a given dimension indicates they use that dimension less, it may be because they have a different understanding of the associated label that, were it encoded some other way, they would prioritize more.

For example, consider figure Figure 3.4. The similarity dimensions are ordered by strength of usage by the population at large. Recall that the Hebart et al. dimensions are ordered by the sum of their scores for all objects; in other words, the dimensions are ordered by how important they are for relating objects for the adult embedding-training population as a whole.

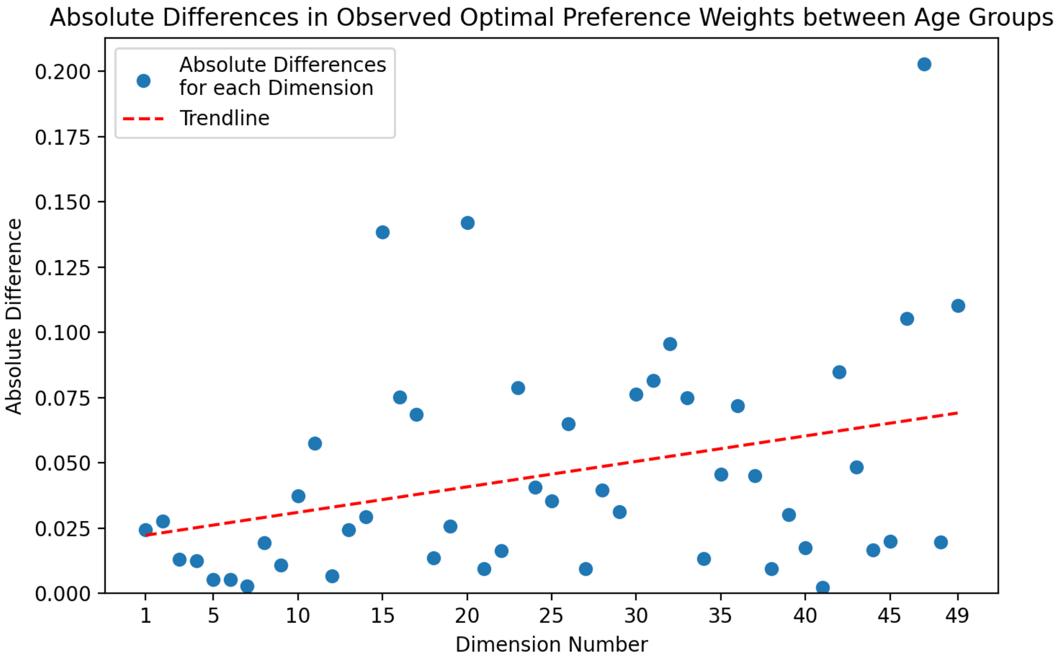


Figure 3.4: The absolute differences between the all-response optimal reweighting preference weights for ages 25–35 and ages 50–60 (results for ages 50–60 minus results for ages 25–35). The x-axis gives the likeness dimension, while the y-axis gives the absolute value of the observed difference between age groups in preference for that dimension. The preferences here were learned over all responses from each age group.

The increasing trend in absolute differences can thus be interpreted in two ways:

1. Firstly, this trend could be because adults, by and large, do compare objects primarily along similar lines regardless of whether they are in their mid-20s-to-early-30s or their fifties. Under this assumption, it would be unsurprising that the most salient dimensions of the original embedding remained more consistent in usage across age groups.

2. Secondly, this trend could be because both groups comprise large portions of the original embedding's training population. The dimensions reflect training-population-wide similarity-judgment decisions, and so the dimensions most relevant to holistic model performance could consequently be ones that describe decisions well for both of our adult age groups.

Under either interpretation, Dimensions 11, 15, and 20 offer particular interest, as the differences in optimal preference weights are both furthest from the red trendline and significant. Dimension 11 has the label “colorful,” Dimension 15 has the label “disc-shaped/round,” and Dimension 20 is “long/thin.” All three dimensions explained older adults’ responses more so than younger adults’, and in all three cases, the dimensions are heavily perceptual in nature, having little to do with functionality, context, or make. Based on these results, older adults may rely more on these perceptual features when judging object similarity than younger adults.

Figure 3.5 offers a more nuanced look at the trends for each age group.

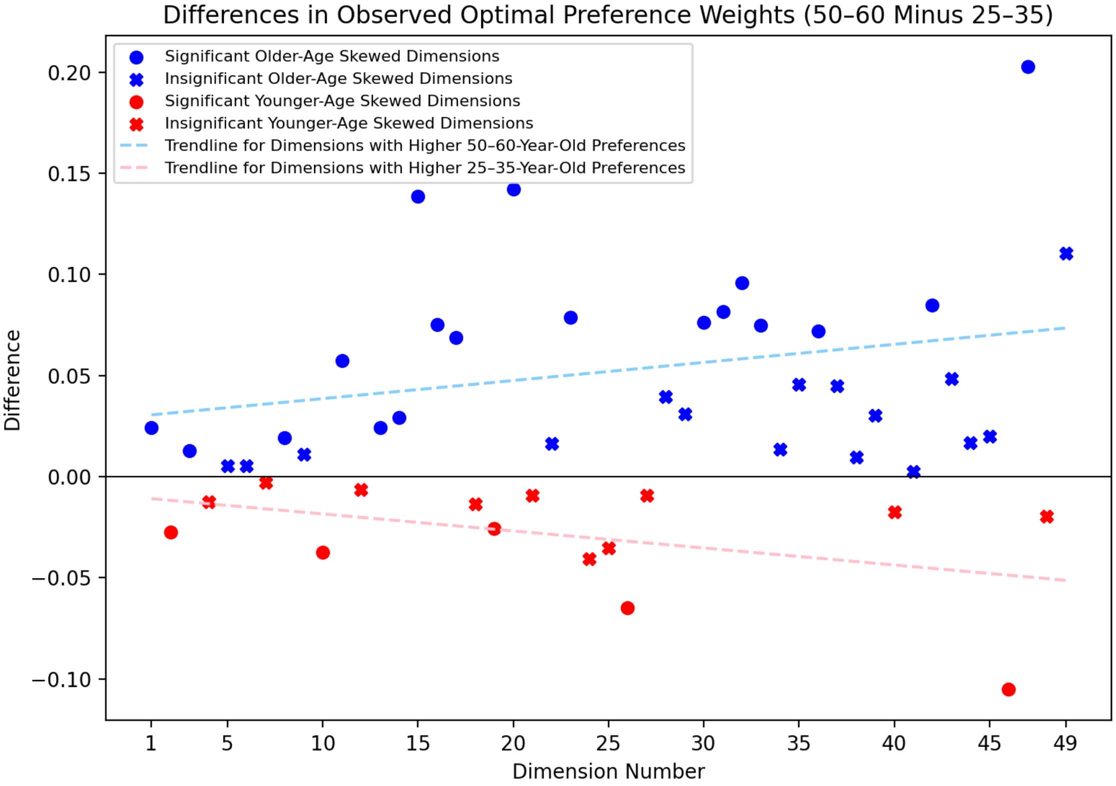
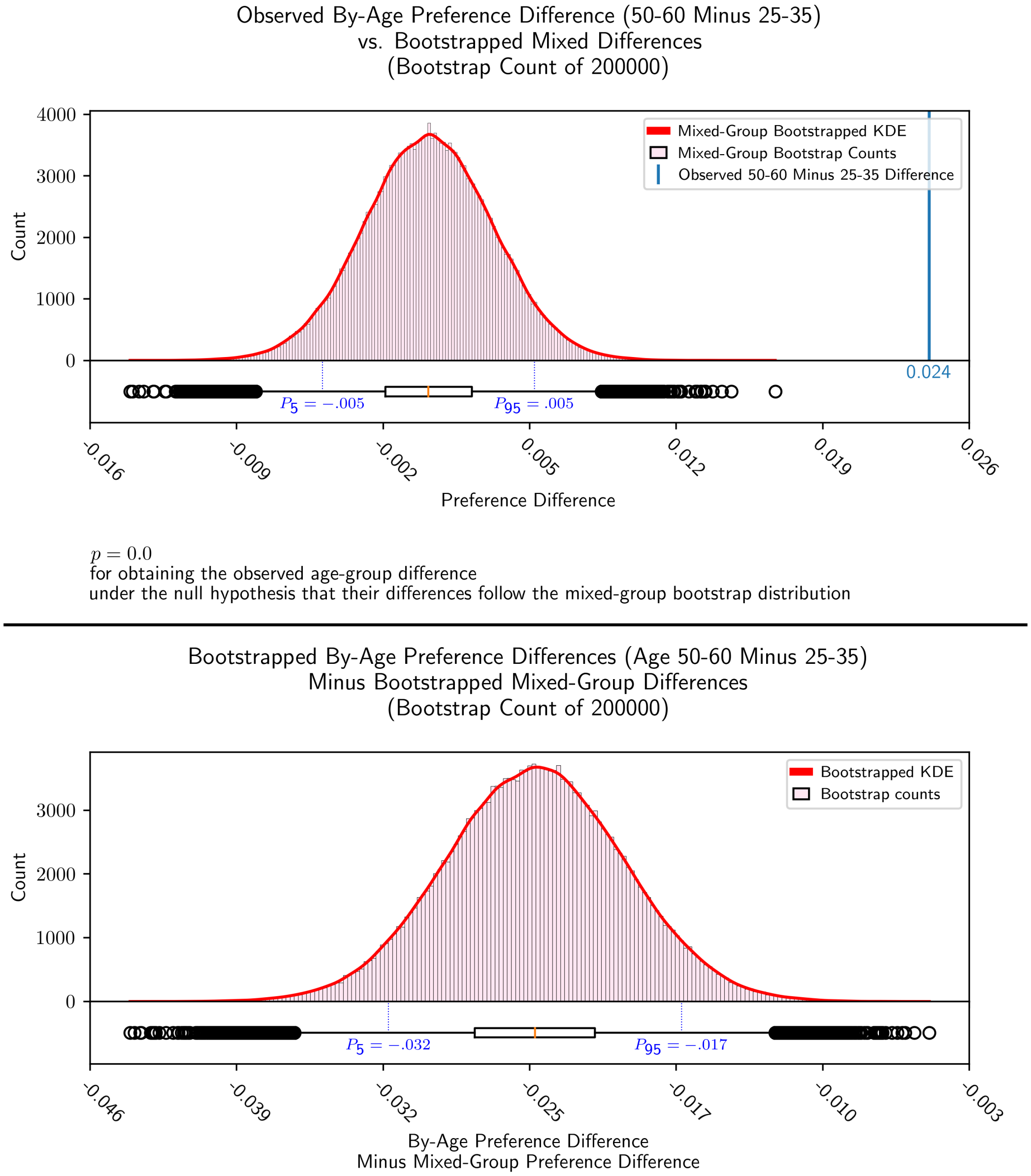


Figure 3.5: The differences between the all-response optimal reweighting preference weights for ages 25–35 and ages 50–60 (results for ages 50–60 minus results for ages 25–35). The x-axis gives the likeness dimension, while the y-axis gives the observed difference between age groups in preference for that dimension; the preferences here were learned over all responses from each age group.

The other largest difference from the trendline that we observe is for Dimension 46, “medicine-related,” with a disproportionate usage by younger adults over older adults. It seems unlikely that the immediate explanation that younger adults focus more on medicine than older adults is plausible. Instead, we suspect that older adults may have a more nuanced, or at least differentiated, view of medicine-related features than do younger adults. For instance, they may not consider medicine-related objects as being as readily similar due to likely having had more interactions with them, whereas younger adults presumably have had less experience with them.

The bootstrapped age-preference plots, preference-difference distributions, confidence intervals, and test results for each age group are given in Appendix A. For illustrative purposes, one such of each have been included below in Figure 3.6 and Figure 3.7.

Dimension 1: made of metal/artificial/hard



$p = 0.0$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

$\mathrm{CI}_{90} = [-0.032, -0.017]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

Figure 3.6: Plots of the bootstrap test and confidence interval (details in Section 3.4.8.1 and Section 3.4.8.2) for Dimension 1, provided here for exemplary purposes. The upper graph tests the probability of observing the full-age-group preference difference (blue line) under the hypothesis that it was sampled from the distribution of randomized groups' preference differences (pink/red distribution). The x-axis gives the difference in preference weights between each age group. Thus, the blue line lying right of zero indicates a greater preference by 50–60-year-olds for the Dimension 1 than by 25–35-year-olds. It being outside the pink distribution connotes significance. The lower graph is a correlated (but different) confidence interval. All dimensions are found in Section A.1.

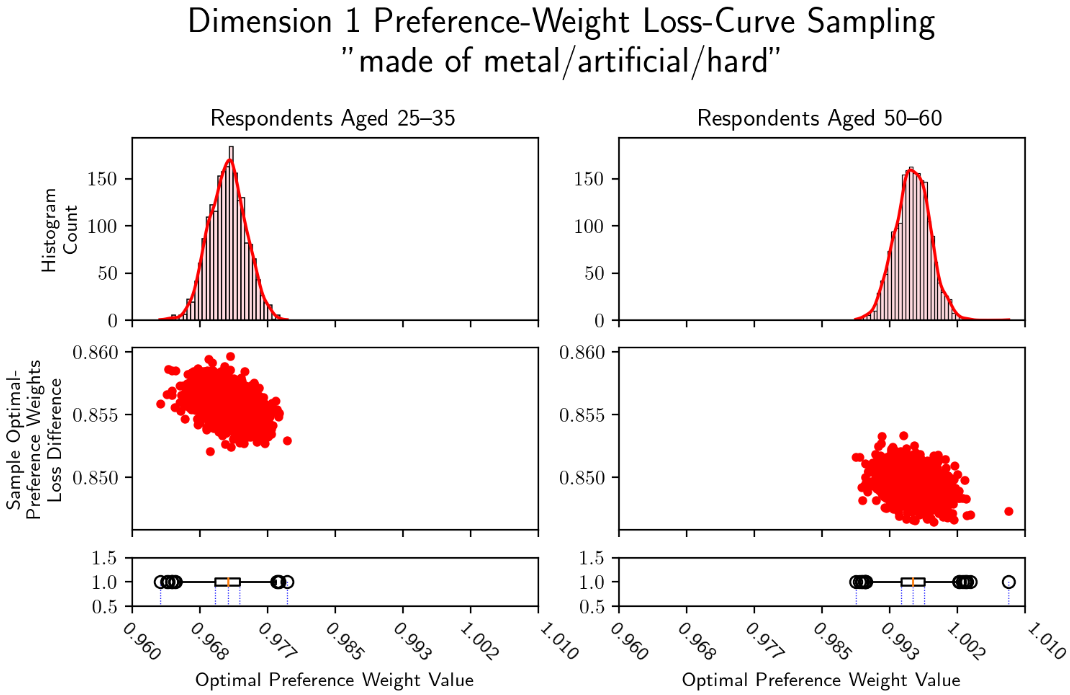


Figure 3.7: Plots comparing the distributions of optimal likeness-preference weights for each age group for Dimension 1, provided here for exemplary purposes. The bootstrapped preferences for the 25–35-year-olds' responses are on the left, while 50–60-year-olds' are on the right. The x-axis gives likeness-preference-weight value, and both distributions are on the same horizontal scale. Results for all dimensions can be found in Section A.2. A lack of overlap between the two age groups' preference-weight distributions is correlated with significant differences between each age group's preferences, but these distributions should not be directly used for statistical testing due to some level of difference being explained by stochasticity. See Section A.1 for distributions valid for statistical significance and Section 3.4.8.2 for an explanation of those distributions.

3.6 Chapter Appendix: Cross-Entropy-Loss Gradient

The derivative of the cross-entropy loss with respect to the rescaling weight at dimension p, w_p is given by

$$
\begin{aligned}
&\frac{\partial}{\partial w_d} \sum_{(i,j,k)\in\mathcal{S}} H(q,p)_{\substack{\text{object set is } \{i,j,k\},\\ \text{k is the odd-one-out}}} \\
&\qquad\qquad = \frac{\partial}{\partial w_d} \sum_{(i,j,k)\in\mathcal{S}} - \ln \frac{e^{\vec{x}_i\vec{x}_j}}{e^{\vec{x}_i\vec{x}_j} + e^{\vec{x}_i\vec{x}_k} + e^{\vec{x}_j\vec{x}_k}} \\
&\qquad\qquad = \sum_{(i,j,k)\in\mathcal{S}} -2w_d \vec{x}_{i0}^{(d)} \vec{x}_{j0}^{(d)} - \frac{e^{\vec{x}_i\vec{x}_j}\vec{x}_{i0}^{(d)}\vec{x}_{j0}^{(d)} + e^{\vec{x}_i\vec{x}_k}\vec{x}_{i0}^{(d)}\vec{x}_{k0}^{(d)} + e^{\vec{x}_j\vec{x}_k}\vec{x}_{j0}^{(d)}\vec{x}_{k0}^{(d)}}{e^{\vec{x}_i\vec{x}_j} + e^{\vec{x}_i\vec{x}_k} + e^{\vec{x}_j\vec{x}_k}}
\end{aligned}
\tag{3.9}
$$

where

- H is the cross-entropy loss function
- i, j, k denote the three objects of a triplet, where k is the true odd-one-out
- z_c where $c \in \{i, j, k\}$ represents the similarity between the pair $\{i, j, k\} \smallsetminus \{c\}$ (see Equation 3.5)
- $\mathbf{z} = \{z_i, z_j, z_k\}$
- σ is the softmax function (see Equation 3.5)
- q is the probability of an object being the odd one out (so 100% for the human-labelled odd-one-out, 0% for any other object)
- p is the estimated probability the model gives that a given object is the odd-one-out
- x_c is the reweighted embedding vector for object c; i.e. $\left[x_{c0}^{(1)}, x_{c0}^{(2)}, ..., w_d x_{c0}^{(d)}, ..., x_{c0}^{(49)}\right]$, where x_{c0} is the unreweighted embedding vector for object c, and d is the unique dimension being reweighted (as per Section 3.4.7.2)

Chapter 4

Differences in Taxonomic- and Thematic-Relation Preferences between Children and Adults

4.1 Overview

As people age, they adjust how they associate different objects. Here, we experimentally examine differences between child and adult prioritization of taxonomic and thematic features in performing an odd-one-out-among-three object comparison task. We leverage a prior odd-one-out adult dataset and associated response-predicting computational model, as well as an age-of-acquisition dataset. With these responses and the model, we derive sets of three objects with known adult odd-one-out responses and scores indicating how much taxonomic/thematic features hampered those responses. We administer these triplets as part of an odd-one-out study to children of age 6, then compare the calculated taxonomic/thematic scores with whether the children deviated from the original adult response. Corroborating previous work, we examine the effect of age on taxonomic- and thematic-relation preference, where taxonomic and thematic preferences are measured relative to each other. Novelly, we also show how to use the computed taxonomic and thematic scores to control for relationship strength and measure taxonomic and thematic preferences independently of one another.

4.2 Questions Examined

Here, we seek to answer the question of how age impacts taxonomic and thematic relation preference in comparing objects. To address this, we look at the following questions and ways of answering them:

Question “Age vs. relative preference for taxonomic and thematic relations”: do children (age 6) exhibit stronger taxonomic- or thematic-relation preferences than adults, where taxonomic- and thematic-relation preferences are measured relative to each other?

Procedure Using pairs of object triplets annotated with taxonomic- and thematic-relation scores and known adult odd-one-out responses for these triplets, collect child responses for the same odd-one-out task. Then, perform pairwise tests to determine if either type of relation explains any resulting child–adult response discrepancies significantly more than the other.

Question “Age vs. absolute preference for taxonomic and thematic relations”: do children (age 6) exhibit stronger taxonomic- or thematic-relation preferences than adults, where taxonomic- and thematic-relation preferences are measured independently from each other?

Procedure Using object triplets with varying taxonomic and thematic scores and known adult responses, collect child (age 6) responses. Then, perform regression to determine if the strength of either relation is significantly correlated with any resulting child–adult response discrepancies.

4.3 Introduction

How do young children contrast with adults in the traits they use when evaluating two objects' similarity?

Comparing objects is a multifaceted task. Thus, when a person is asked to judge how similar two objects are, they do so along many paradigms. For instance, "dogs" and "bones" are similar in that we commonly encounter them with each other, but they are not similar in appearance or function. On the other hand, we do not often encounter "dogs" and "bears" in the same setting. Instead, we recognize them as similar by virtue of shared features such as legs and fur.

Ways such as these of comparing objects are called semantic relations. The two broad classes of semantic relations alluded to in the "dog–bone" and "dog–bear" examples are called thematic relations and taxonomic relations. A thematic relation between two objects is a semantic relation wherein those objects belong to a theme or co-occur in some context. In this manner, a thematic relation between "dog" and "bone" is "pet-related." Meanwhile, a taxonomic relation between two objects is a semantic relation wherein those objects share features or functions. For instance, many "dog–bear" relations are taxonomic, such as "are-mammals" or "have-legs."

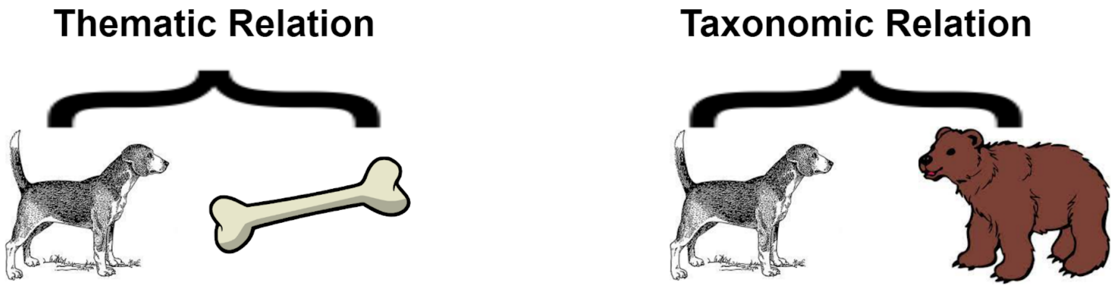


Figure 4.1: Example thematic and taxonomic relations. "Dog" and "bear" are taxonomically related because of shared features, or traits that could be generated by looking at each individually, such as "has-paws" or "has-fur." "Dog" and "bone," meanwhile, are thematically related because they co-occur in contexts like a pet store.

Individuals vary on how much they prioritize different relations when determining two objects' similarity, but across larger groups, certain trends emerge. Consider the connection between age and the preference for the taxonomic and thematic classes of relations described earlier. Prior work [10] shows that, when compared with adults, young children exhibit a stronger affinity for thematic (co-occurring) relations than taxonomic (shared-feature) relations. For example, one study [9] found a thematic preference for children younger than age six that becomes increasingly taxonomic by age ten. This taxonomic preference continues through a person's early twenties, then reverts back to a thematic preference by a person's mid-sixties. See Section 2.2 for more details.

A common form of testing this is to select a base object ("dog") and then present it alongside a taxonomically related object ("bear") and a thematically related object ("bone"). A set of three objects such as this is known as a ***triplet***. Selections of three objects are presented like this to people of different ages, who are then asked to indicate which object is most similar to the base object. Differences in responses across different age groups indicate changes in taxonomic/thematic preferences.

These experiments are useful for determining shifts in preferences with age. However, under most setups, the questions are agnostic to the varied strengths of taxonomic and thematic relations. For example, the taxonomic relation "mammal-related" applies to both "dog–bear" and "dog–dolphin" comparisons, but the strength between "dog" and "bear" is stronger due to elements like both having hair. Experiments without this consideration can only tell us about the strengths of taxonomic and thematic relation preferences relative to one another, and they cannot easily tell us about absolute trends and cannot tell us about changes in behavior as these strengths vary.

Unfortunately, accommodating for relations' strengths is nontrivial. Doing so requires numeric measures of the those strengths. To account for the strength of the broad classes of taxonomic relations and thematic relations on the whole, one needs numeric measures that characterize those types of relations. Fortunately, recent advances in modelling object comparison give us quantitative measures of more specific relations, which we can use to derive these numeric taxonomic and thematic measures.

In this chapter, to address the question, “how do young children compare with adults in the traits they use when evaluating two objects’ similarity?” we set up two experiments:

1. We examine the effect of age on taxonomic- and thematic-relation preference, where taxonomic and thematic preferences are measured relative to each other. We do this for different levels of expected taxonomic and thematic strengths.

2. We examine the effect of age on taxonomic- and thematic-relation preferences independently of each other. We do this for different levels of expected taxonomic and thematic strengths.

We use prior adult responses to an object-similarity task and an associated computational model to select triplets of three objects for administration to children. We present these object triplets to children and ask them, “which of these three objects is the odd one out?” This questioning gives us a set of child responses for usage alongside the existing adult responses. We use measures for the respective strengths of taxonomic and thematic relations to derive pairs of triplets along taxonomic and thematic lines. We also use these measures to ensure a variety of taxonomic relations, thematic relations, and relation strengths of each are present. We administer these triplets, paired along taxonomic and thematic lines as we set them up to be, to six-year-old children. We interpret differences between the resulting children’s responses and the original adult responses as changes in taxonomic or thematic preference.

4.4 Methodology

4.4.1 Overview

Our overarching goal is to investigate the interaction between age and type-of-relation preferences in comparing objects. In particular, we focus on two age groups: six-year-olds and adults, and two types of object-comparison relations: taxonomic and thematic. Our study differentiates itself from prior work in two ways. The first point of differentiation is computationally deriving the object-comparison questions rather than manually deriving them. The second point of differentiation is using numeric scores to control for the strengths of each taxonomic and thematic similarity rather than treating all taxonomic (or, separately, thematic) similarities as interchangeable in this regard. This second point allows us to look at taxonomic and thematic preferences individually and quantitatively, which prior studies were unable to do.

We perform two experiments:

1. We examine the effect of age on taxonomic- and thematic-relation preferences, where taxonomic and thematic preferences are measured relative to each other. We do this for different levels of expected taxonomic and thematic strengths.

2. We examine the effect of age on taxonomic- and thematic-relation preferences independently of each other. We do this for different levels of expected taxonomic and thematic strengths.

At a high level, our approach is as follows. At the outset, we have a set of adult responses to an odd-one-out-among-three object-comparison task. Each response is to a set of three objects, or object triplet, that defines the odd-one-out-question at hand. We give these questions scores that indicate the extent that the deviation of someone's response from the original adult data would signal a preference for taxonomic or thematic relations. We select a small subset of these questions based on these scores. We then ask six-year-old children these questions, recording the situations where their responses deviate from the adults'. The correspondences between these deviations and the taxonomic and thematic measures reveal differences by age in type-of-relation prioritization. Below, we give an overview of the elements of our procedure in five parts:

1. The prior dataset
2. How we generate questions appropriate for six-year-olds
3. How we generate scores for testing similarity preferences
4. How we use these scores to generate object triplets
5. How we use these triplets to form and test hypotheses

First, we discuss the prior dataset. We have an existing object-comparison answer dataset to work with [13]. This dataset contains millions of adult responses to an object-comparison task, the odd-one-out task. This task, which we use in our experiment, involves presenting three object images at a time and asking respondents to identify the least similar among the three. Object images are taken from the THINGS initiative [14] [15], which provides a dataset of object images and associated metadata, including English-word annotations.

Secondly, we discuss generating age-appropriate questions. As we have prior adult data, we need only collect odd-one-out-task responses from six-year-olds. We take a child's response for a set of three object images, then compare that to existing adult responses for those objects. To ensure children can reason with these object images, we must make certain that a given six-year-old child understands those objects conceptually. Fortunately, all objects under consideration are part of the THINGS initiative, so each object has an associated English label. We treat the age of learning the word, otherwise known as the word's age of acquisition, as a proxy for the age of understanding of the underlying concept. Then, we use existing age-of-acquisition measures [26] to assign ages at which the THINGS initiative's objects would be understood accordingly. Finally, under these new annotations, we filter the objects to those understandable at or below age six.

Thirdly, we discuss generating numeric taxonomic and thematic scores. In order to generate these scores for the triplets, we use the dimensions of the prior odd-one-out-predicting model trained on the adult dataset. The model has 49 dimensions, entailing 49 values for each object. Notably, the meanings of these values are consistent and human-identifiable: one value is the "metallicity" of the object in question, another value is "body-part-relatedness," and so on. These values refer to how much the model uses that aspect of relatedness of the object for general comparison. In other words, each value refers to the general strength of a specific type of relation, such as "metallicity" when other objects are compared with the object. We give these relations taxonomic and thematic scores by means of a survey.

Continuing this third goal, we then generate two scores for each triplet: one taxonomic, one thematic. These scores depend only on the model's dimensions (which have been labelled as taxonomic or thematic, as described above) and the adult odd-one-out response. Intuitively, each score denotes how significantly some taxonomic or thematic factor influences the model away from the original adult response and toward some alternative. For this reason, we will henceforth refer to such a taxonomic or thematic score for a triplet equipped with an odd-one-out response as a triplet's ***taxonomic or thematic confusion***. Without loss of generality with respect to taxonomicity and thematicity, we will explain exactly what we mean by taxonomic confusion. A ***dimension confusion score*** for a triplet is that dimension's influence on the computed similarity of the odd-one-out object with one of the non-odd-one-out objects. The taxonomic confusion for a triplet is the maximum among these values for the taxonomic dimensions. We treat the model as a proxy for adult thinking, so higher taxonomic scores signify an increased likelihood of an adult adjusting their response if they had higher prioritization of taxonomic relationships.

For the fourth element of our procedure, we identify pairs of ***taxonomically and thematically confusing object*** triplets. By this, we mean pairs of triplets where one triplet's associated taxonomic confusion score and another triplet's associated thematic confusion score are close to each other. We do this for diverse taxonomic/thematic scores, producing many such pairs. We additionally control for a few other factors (explained in further detail in Section 4.4.6) when forming these pairs.

Fifthly and finally, we discuss how these combine into testable experiments. For both experiments, we administer the triplets to children aged 6 as part of the odd-one-out task and record whether a child's response matched the adult's response in the prior dataset; these are ***response discrepancies***. Next:

- For the first experiment, we want to examine the effect of age on taxonomic- and thematic-relation preferences, where taxonomic and thematic preferences are measured relative to each other. To do this, we consider the pairs of taxonomically and thematically confusing triplets. We perform a paired statistical test on the response discrepancies to determine whether children's responses deviated from adults more or less often for taxonomically confusing triplets than thematically confusing triplets.

- For the second experiment, we want to examine the effect of age on taxonomic- and thematic-relation preferences independently of each other. To do this, without loss of generality, we consider the taxonomically confusing triplets at various amounts of taxonomic confusion. We perform regression analysis on the response discrepancies to examine how children prioritize taxonomic thinking as compared to adults. We then repeat these two steps for thematically confusing triplets.

Below, we go into more detail about the various elements of our project setup.

4.4.2 Age and Animacy Filtering

When choosing triplets of three objects to administer to children, we first filter these triplets based on two features inherent to the objects themselves: the age at which the images are understandable and animacy. We perform age filtering to ensure that six-year-olds can understand the images presented to them. As part of age filtering, we also remove any triplets involving the objects “gun,” “shell” (`shell1` in the dataset, which is in the weaponry sense), “ashtray,” “helmet,” “sticker,” “uniform,” “spider,” “spider web,” and “cross” (which includes the Iron Cross) as a precaution to avoid potential fright, offence, or general inappropriateness.

We perform animacy filtering to ensure that a majority of the results are not explainable by two objects being humans/animals and the third object not falling into that situation (or vice-versa).

4.4.3 Ages of Acquisition

Because we are using the objects of the THINGS dataset and the prior adult responses of Hebart et al.’s study to survey six-year-olds, we need to restrict the triplets administrable to the children to those that would be understandable by them. The THINGS dataset has WordNet metadata, which we use to obtain estimates on the ages of acquisition of the objects in the dataset and restrict our consideration to those ostensibly understandable by most people six or over.

For every object in the THINGS dataset, we take its annotated label and that label’s synonyms. For those entries with annotated labels in the age-of-acquisition dataset, we simply take the given age of acquisition of that annotated label as a ***canonical age of acquisition*** for the object. For those that did not, we assign them an infinite score for this value.

In order to alleviate issues presented by homonyms (baseball “bat” vs animal “bat”), we additionally take the ages of acquisition of all synonyms present in the age-of-acquisition dataset. We do not assign unknown words the value of infinity here— in order to calculate an ***alternative age of acquisition*** for a THINGS object, we take the maximum of the ages of acquisition of all of the synonyms. This is because requiring objects to have only known synonyms would result in too much of a loss of otherwise viable objects.

Finally, we filter the triplets dataset to only those triplets where all objects involved had a canonical age of acquisition and an alternative age of acquisition below six years. In using the maximum age between the canonical age of acquisition and the alternative age of acquisition, we enforce a more conservative estimate on the age of acquisition. This setup helps ensure that the actual initial age at which all of the objects in the filtered triplets are understood in common is not higher than six.

4.4.4 Animacy Filtering

One particularly strong indicator of object-relatedness is ***animacy***. Animacy, here, is a measure of whether something is treated as a human/animal or treated as if they were sentient. Generating questions without accounting for this resulted in a large number of triplets where two were taxonomic and one was thematic, or where two were thematic and one was taxonomic. The ostensible explanation for a resulting response difference with these triplets would be that animacy was the driving factor.

To ensure we have either exclusively animate or exclusively inanimate sets of three objects, we label all 1854 objects in the THINGS initiative as either animate or inanimate[1]. This labelling is facilitated by the Google Translate Python API, which we use to determine Russian equivalents of the canonical words given by the existing THINGS annotations. We then leverage the declensions of those translations, which in Russian differ depending on the animacy of the object being referred to. We choose to use a measure of real-world linguistic animacy rather than merely labelling objects

[1] There are (unpublished at time of writing) animacy labels for the THINGS initiative [27] that we were not aware of at the time of working on this portion of the project.

based on human/animal features under the expectation that natural animacy would appeal to early-age world conceptualization more (as an example of where Russian differed from this simpler approach, “snowman” is considered animate). Our choice of language is decided by the availability of quality labels; many Slavic languages display animacy as a grammatical feature, but Russian has one of the most sizable online corpora to draw from. In this case, we use the English [28] and Russian [29] Wiktionary sites).

Words with ambiguous labels—that is, those where the declension for the sense of the noun in question could be inanimate or inanimate, or those where the animacy was rooted in linguistic history (like the word for kite being the word for serpent)—are annotated with WordNet [18]. For these entries, if WordNet has an associated “animal” or “human” tag, the object in question is considered animate; else, it is considered inanimate.

After we determine these labels, we filter the dataset of adult odd-one-out questions and responses down to those where the questions contain three animate or three inanimate objects.

4.4.5 Triplet Numeric Scoring

4.4.5.1 Summary

Two goals for our project are computationally deriving the object-comparison questions, rather than manually deriving them, and controlling for the strengths of each taxonomic and thematic similarity, rather than treating all taxonomic (or, separately, thematic) similarities as interchangeable. To these ends, we obtain what we call measures of taxonomic and thematic confusion. Given a set of three objects and a response, taxonomic and thematic confusion signify the expected willingness of a person to deviate from that response due to prioritizing taxonomic or thematic relations more than the original respondent. We obtain these scores by manipulating the embeddings of an odd-one-out predicting model trained on a prior adult response dataset.

4.4.5.2 Taxonomic and Thematic Dimension Determination

Eleven professors and graduate students in computing science and in psychology were given explanations of taxonomic and thematic relationships and asked to identify each of the model's dimensions as "taxonomic," "thematic," or "unknown." Survey results can be found in Appendix C, while the survey itself can be found in Appendix B. Dimensions where either taxonomic or thematic labels outnumbered the other by more than four were considered to be indicative of that relationship type, resulting in 13 taxonomic and 11 thematic dimensions. All 13 dimensions considered "taxonomic" were incidentally considered "taxonomic" by a majority of respondents, and all 11 dimensions considered "thematic" were incidentally also considered "thematic" by a majority of respondents. Neither of these characterizations was a requirement for producing the respective dimension labels; nor was a majority response a sufficient condition for producing the respective dimension labels in practice (due to the allowance of "unknown" responses).

4.4.5.3 Overall/Taxonomic/Thematic Confusion and Taxonomic/Thematic Affirmation Scores

We wish to determine whether young children value taxonomic and thematic relationships relatively more so or less so in comparison with adults. To do this, we want some measure by which to select triplets that, given a child's difference in response from the adult, would indicate that the child were or were not prioritizing these relationships. We accomplish this using measures of ***taxonomic confusion*** and ***thematic confusion***. Given a set of three objects and a response, taxonomic and thematic confusion signify the expected willingness of a person to deviate from that response due to prioritizing taxonomic or thematic relations more than the original respondent. We also consider ***overall confusion***, a similar concept but instead accounting for all model dimensions, and ***taxonomic and thematic affirmation***, a measure of how much taxonomic or thematic relations contribute to the associated adult decision.

4.4.5.4 Taxonomic/Thematic Confusion

The Hebart et al. model computes the similarity between two objects by taking the internal 49-dimensional vector representations of those two objects, v_1 and v_2, multiplying the elements of each element-wise, then summing the resulting products ($v_1 \cdot v_2$). Each of those products indicates the model's usage of the originating dimension in determining the similarity of the objects. By examining the taxonomic dimensions' products (***taxonomic products***) and the thematic dimensions' products (***thematic products***), then, we learn how much the model prioritized taxonomic and thematic features in its computation. By treating the model as a proxy for adult human thinking, this gives us numerical estimates to explain how much the adult prioritized taxonomic and thematic features in their decision-making.

We consider the taxonomic/thematic products and determine the size of the largest taxonomic product of the similarities that *did not* yield the correct human odd-one-out. This is the ***taxonomic confusion***. Likewise, we determine the size of the largest thematic product of the similarities that did not yield the correct odd-one-out. This is the ***thematic confusion***. The intuitive understanding of this is that the higher these values are, the more a single taxonomic/thematic dimension steers someone answering the odd-one-out task away from the answer under consideration. It should be noted that these values depend on an initial existing human choice of odd-one-out. We choose the highest among the confusion values rather than the median or mean. We control for these values in Section 4.4.6.

4.4.5.5 Overall Confusion

In a given triplet, the model has similarity judgments between each pair of objects. It chooses the pair with the highest similarity s_1, but were that value lower, it would choose the pair with the next highest similarity s_2. Suppose the model accurately represents people's object-similarity preferences, but that humans exhibit more variability. As the difference between s_1 and s_2 becomes smaller, the chances of a human responding differently than the model increase. We thus call this difference ***overall relative confusion*** (or ***overall confusion***).

We want to reduce disparate impacts of overall relative confusion among our triplets. Consequently, we control for this value in Section 4.4.6.

4.4.5.6 Taxonomic/Thematic Affirmation

The opposite of taxonomic and thematic confusion we describe as ***taxonomic and thematic affirmation***. These are the maximum taxonomic products of the non-odd-one-out pair and the maximum thematic products of the non-odd-one-out pair. Intuitively, these are measures of how much the taxonomic and thematic relations facilitated decision-making. We control for these values in Section 4.4.6.

4.4.6 Triplet Selection and Pairing

4.4.6.1 Summary

We wish to select pairs of triplets where the following is true: first, one triplet has a similar level of thematic confusion as the other has taxonomic confusion; and second, these levels of taxonomic/thematic confusion are varied. Doing so lets us administer these triplets to children, interpret adult-child response differences as taxonomic/thematic preference differences, and directly compare the rates of these differences across each type of relation. We control for a few additional factors as well.

4.4.6.2 Taxonomic/Thematic Candidate Dataset Splitting

We consider pairs of triplets such that a difference between the adult and child responses for one of the triplets indicates an increase in taxonomic thinking (a ***taxonomically confusing triplet***), while the same such difference results in an increase in thematic thinking for the other (a ***thematically confusing triplet***). To accomplish this, we regard the dataset twice: once for selecting thematically confusing triplets and then again for selecting taxonomically confusing triplets. Thus, we duplicate the dataset into two copies—a set of 26 454 candidate taxonomically confusing triplets and a set of 26 454 candidate thematically confusing triplets. Because we want to pair them for certain properties, we draw from each copy simultaneously to account for those properties.

4.4.6.3 Initial Filtering

For both sets of triplets, we calculate the 80th percentile of the thematic affirmation, taxonomic affirmation, and (additive) inverse total relative confusion scores among all triplets. We also calculate the 80th percentile of thematic confusion for the candidate taxonomically confusing triplet set and the 80th percentile of taxonomic confusion for the candidate thematically confusing triplet set. We cull any triplets with values below these percentiles (or below that percentile for thematic and taxonomic confusion, respectively, for the candidate taxonomically confusing triplets and candidate thematically confusing triplets). This gives us a set of 11 351 candidate taxonomically confusing triplets and a set of 10 431 candidate thematically confusing triplets.

4.4.6.4 Generating Pairs of Triplets with Different Values of Taxonomic and Thematic Confusion

While we want pairs of triplets where one triplet has a similar taxonomic confusion to the other's thematic confusion—i.e., both values are within some distance from a mean value—we want to make sure that we have a variety of triplets with different mean values. As such, we take pairs of triplets where the taxonomically confusing triplet's taxonomic confusion and thematically confusing triplet's thematic confusion are within a given range for different ranges. The specific ranges we consider are $[0, 0.1), [0.1, 0.2), \ldots, [1.1, 1.2),$ and $[1.2, \inf)$. We select these ranges due to the distribution of the set of triplets we have to work with, as we need to ensure we have sufficient numbers of triplets to work with when balancing for other important factors. These factors are explained in the next section. We split the candidate triplets into 12 sets based on these ranges, then select from those sets. It is worth noting at this point that the eleventh and twelfth sets ($[1.1, 1.2), [1.2, \inf)$) have fewer triplets than the rest due to the initial filtering in Section 4.4.6.3; this means that those sets will end up with fewer triplets than the rest.

4.4.6.5 Generating Pairs of Triplets with Similar Taxonomic and Thematic Confusion

We want to ensure taxonomically confusing triplets and thematically confusing triplets have similar values of (respectively) taxonomic confusion and thematic confusion. However, we simultaneously want to choose triplets that have the lowest possible amounts of taxonomic affirmation and thematic affirmation. Were these values unaccounted for, child–adult response discrepancies could instead be interpreted in terms of these affirmation values. We also want to choose pairs of triplets where each member of the pair has a similar overall relative confusion value, as we expect overall relative confusion to be a strong measure of whether someone would have a different response from the original adult across all reasons tracked by the model.

For this reason, we generate vectors from these values. Specifically, we generate two vectors, one for each triplet, as given by Equation 4.1 and Equation 4.2.

$$\begin{gathered} v_{\text{taxonomically confusing triplet}} = \\ (\text{taxonomic confusion} \cdot 8, \text{thematic confusion} \qquad (4.1) \\ \text{taxonomic affirmation}, \text{thematic affirmation}, \text{overall relative confusion}) \end{gathered}$$

$$\begin{gathered} v_{\text{thematically confusing triplet}} = \\ (\text{thematic confusion} \cdot 8, \text{taxonomic confusion}, \qquad (4.2) \\ \text{thematic affirmation}, \text{taxonomic affirmation}, \text{overall relative confusion}) \end{gathered}$$

We multiply the taxonomic and thematic confusion by 8 because we consider controlling for them particularly important compared to the other influences. We additionally avoid consideration of any pair of triplets where both triplets contain the same objects.

For every pair of triplets in the dataset, we calculate the $\mathcal{L}_2$ distance between them under these vectors. For each of the ranges of values discussed in Section 4.4.6.5, we take the top 320 pairs (if possible, if there are fewer available we take that many) of triplets, which we will filter down to 20 (or fewer, if there are not 20 to be had) in the next section.

4.4.6.6 Diversifying Triplets

For any given taxonomic confusion score, thematic confusion score, taxonomic affirmation score, and thematic affirmation score, there is an associated dimension that gives that score (recall that for all but overall confusion we take the maximum of the relevant dimensional products, and so one of those dimensions yields a given score). We want to make sure the resulting triplets cover a wide variety of these dimensions. As such, we perform two filtering steps: one less computationally intense to bring the number of triplet pairs down to 32, and then one more computationally intense to bring that number down to 20.

Weaker Triplet Diversification

Without loss of generality, consider one of our bins. We aim to bring our triplets from 320 (or fewer) down to 32 (or fewer).

- We start with a list of candidate pairs of triplets, $\text{pairs}_{\text{candidate}}$.
- We keep a running list of pairs of triplets, $\text{pairs}_{\text{accepted}}$. This starts out empty and builds up until we hit 32 triplets.
- We keep track of the (non-unique) taxonomically confusing and taxonomically affirming dimensions, $\text{dims}_{\text{tax. conf.}}$, and $\text{dims}_{\text{tax. aff.}}$, among the taxonomic triplets within $\text{pairs}_{\text{accepted}}$. Here, the taxonomically confusing/affirming dimension means the dimension whose product yielded the taxonomic confusion/affirmation for the triplet.
- We keep track of the (non-unique) thematically confusing and thematically affirming dimensions, $\text{dims}_{\text{them. conf.}}$, and $\text{dims}_{\text{them. aff.}}$, among the thematic triplets within $\text{pairs}_{\text{accepted}}$. Here, a given thematically confusing/affirming dimension means the dimension whose product yielded the thematic confusion/affirmation for the triplet.
- We keep track of the (non-unique) maximum dimensionally confusing and maximum dimensionally affirming dimensions, $\text{dims}_{\text{max. tax. dim. conf.}}$, and $\text{dims}_{\text{max. tax. dim. aff.}}$, among the taxonomic triplets within $\text{pairs}_{\text{accepted}}$. We also keep track of the (non-unique) maximum dimensionally confusing and maximum dimensionally affirming dimensions, $\text{dims}_{\text{max. them. dim. conf.}}$, and $\text{dims}_{\text{max. them. dim. aff.}}$, among the thematic triplets within $\text{pairs}_{\text{accepted}}$. Here, a given maximum dimensionally confusing/affirming dimension means the dimension with the highest dimensional confusion/affirmation for a triplet.
- We consider a list of dictionaries $\text{dict}_{\text{counts}}$ that each, when provided a dimen-

sion d, maps d to the number of instances of itself within each of $\text{dims}_{\text{tax. conf.}}$, $\text{dims}_{\text{tax. aff.}}$, $\text{dims}_{\text{them. conf.}}$, $\text{dims}_{\text{them. aff.}}$, $\text{dims}_{\text{max. tax. dim. conf.}}$, $\text{dims}_{\text{max. tax. dim. aff.}}$, $\text{dims}_{\text{max. them. dim. conf.}}$, and $\text{dims}_{\text{max. them. dim. aff.}}$.

We pick a proportion, $p = 0.1$. We perform the following steps 5 times:

1. Consider each pair of triplets triplet $d = (t_{\text{taxonomic}}, t_{\text{thematic}})$ in $\text{pairs}_{\text{candidate}}$.
2. If $\text{len}(\text{pairs}_{\text{accepted}}) = 32$ or $d \in \text{pairs}_{\text{accepted}}$, skip d.
3. Otherwise:
   - 3.1. Consider the taxonomically confusing dimension, taxonomically affirming dimension, maximum dimensional confusion dimension, and maximum dimensional affirmation dimension of $t_{\text{taxonomic}}$, and the thematically confusing dimension, thematically affirming dimension, maximum dimensional confusion dimension, and maximum dimensional affirmation dimension of t_{thematic}.
   - 3.2. Take the count of each of these dimensions within the accepted triplets by applying them to their respective dictionary in $\text{dict}_{\text{counts}}$.
   - 3.3. Turn each of these counts into proportions by dividing them by $\text{len}(\text{pairs}_{\text{accepted}})$.
   - 3.4. If any of these proportions is greater than p, skip d. Otherwise, add d to $\text{pairs}_{\text{accepted}}$.

Finally, we repeat steps 1 through 3 five times each for the proportions $p = 0.11, 0.12, 0.13, \ldots, 0.40$. We pass the resulting list of up to 32 pairs of triplets to the filter in the next section.

Stronger Triplet Diversification

Finally, for our last step of triplet selection, we perform another optimization. We do this to give us a set of 20 triplet pairs where the likeness dimensions producing the highest taxonomic confusion score, thematic confusion score, taxonomic affirmation score, and thematic affirmation score are diverse. We skip this step if the prior step returned less than 20 triplet pairs. In order to achieve this, we generate sets $\{s_i\}$ of 20 pairs of triplets from the list given by the prior step of up to 32 pairs of triplets $\vec{d}$ and consider the following for each s_i:

- the number n_1 of unique taxonomically confusing categories in s_i
- the number n_2 of unique thematically confusing categories in s_i
- the number n_3 of unique taxonomically affirming categories in s_i
- the number n_4 of unique thematically affirming categories in s_i
- every n_i together as the list $\vec{n}$
- the ratio $r = \frac{\bar{n}}{20}$
- the scoring mechanism $\text{score}_{\text{triplet diversity}}(s_i) = \min(\vec{n}) + r$

We compute $\text{score}_{\text{triplet diversity}}(s_i)$ for all possible combinations $\{s_i\}$ of 20 triplet pairs from the starting list of pairs (up to 225 792 840 given a starting list of 32 pairs of triplets) and choose the set s_i with the highest score.

This results in approximately 210 pairs of triplets across the 12 bins. There are not 240 pairs of triplets in total due to later bins having fewer than 20 triplets at the start of this step (see the note in Section 4.4.6.4).

4.5 Triplet Administration

We duplicate the set of 210 pairs of triplets three times for 630 pairs (1260 triplets) total, then shuffle the pairs of triplets.

Triplets were administered to children via our colleagues at the Speech Development Lab at the University of Calgary. Data were collected using the online server-management tool JATOS [30].

The triplets were disseminated amongst 31 participants, all of age 6, under the supervision of their guardian and the researcher at the Speech Development Lab. Each child performed the odd-one-out task for 20 pairs of triplets, taking 15–30 minutes to do so. They were presented with the canonical images present in the THINGS dataset associated with the objects of each triplet.

4.6 Experiment 1: Age and Relative Taxonomic and Thematic Relation Preferences

4.6.1 Experiment Details

We examine the effect of age on taxonomic- and thematic-relation preferences, where the taxonomic and thematic preferences are measured relative to each other. We do this for various amounts of expected taxonomic and thematic strengths.

For the first experiment, we consider whether children think relatively more taxonomic or thematically than adults at different ranges of taxonomic and thematic confusion. Given that we have bins $[0.0, 0.1)$, $[0.1, 0.2)$, $\ldots$, $[1.1, 1.2)$, $[1.2, \text{inf})$ of pairs of taxonomically and thematically confusing triplets organized by level of taxonomic/thematic confusion, for each range of these bins, we aggregate the "do the child and adult responses differ" responses into a contingency table, as in Table 4.1.

We assume there is no difference in the child–adult response discrepancy rate between the thematically confusing triplets and taxonomically confusing triplets. We assume this for each range of taxonomic/thematic confusion. We perform statistical testing on these tables to attempt to disprove this assumption—i.e., to show that either taxonomic or thematic confusion is statistically correlated with children choosing different odd-one-out responses than adults.

More specifically, we choose McNemar's test to avoid issues pertaining to the underlying distributions of the responses. Our samples are random and presumably IID, although there is a possibility that some prompts for a given participant affected their line of thinking for other responses.

Table 4.1: An example contingency table result.

		Thematic Response Difference		
		False	True	
Taxonomic Response Difference	False	39	7	46
	True	1	1	2
		40	8	

4.6.2 Results

The significant results for Experiment 1 are given in Table 4.2.

Full results for Experiment 1 can be found in Appendix D.

Confusion Range	p-value	Statistic	Contingency Table Th. Sw. F T Ta. Sw. F T	Taxonomic Swaps	Thematic Swaps	Significant Thematic Priority?	$p < 0.05$	$p < 0.1$
0.0–0.1	0.077	3.125	F 39 | 7 T 1 | 1	2	8	TRUE	F	T
0.0–0.2	0.034	4.5	F 71 | 14 T 4 | 4	8	18	TRUE	T	T
0.0–0.3	0.091	2.857	F 101 | 23 T 12 | 7	19	30	TRUE	F	T
0.0–1.2	0.093	2.82	F 329 | 74 T 54 | 22	76	96	TRUE	F	T
0.7–0.9	0.072	3.226	F 57 | 21 T 10 | 4	14	25	TRUE	F	T
0.7–1.2	0.074	3.2	F 106 | 29 T 16 | 5	21	34	TRUE	F	T
0.7–10.0	0.093	2.824	F 116 | 32 T 19 | 5	24	37	TRUE	F	T
all	0.101	2.694	F 339 | 77 T 57 | 22	79	99	FALSE	F	F

Table 4.2: Significance testing for relative taxonomic and thematic preferences at various taxonomic and thematic confusion strengths. Only significant and overall results are shown. For all results, see Appendix D.

4.6.3 Discussion

As a recap, in this experiment, we tested for differences between adults (age 18-) and children (age 6) in relative taxonomic and thematic preferences. We did this by pairing triplets where response-switching would be indicative of taxonomic or thematic preferences, the degree to which we referred to as "taxonomic and thematic confusion." Rates of child–adult response-discrepancy for various ranges of taxonomic/thematic confusion, as well as statistical testing to determine if the differences between the rates for taxonomically confusing triplets and thematically confusing triplets are significant, are displayed in Table 4.2. Between pairs, we controlled for overall confusion, taxonomic confusion, thematic confusion, thematic affirmation, and taxonomic affirmation; see Section 4.4.6 for information.

At $p = 0.1$, there are several ranges of values for which there is both enough data and enough of a taxonomic or thematic preference to indicate a statistically significant preference for one or the other, including the range of 0-1.2 that considers nearly all of our full-set data (the ranges 1.1-1.2 and 1.2-inf have fewer triplet pairs than the rest of the triplet sets; see Section 4.4.6.6). In all of these cases, a thematic shift was observed. Given the number of ranges considered, however, even this relatively weak statistical power should be considered with some reservation.

Somewhat confusingly, significant relative thematic preferences by children were observed for lower ranges of taxonomic/thematic confusion (<0.3) and higher ranges (<0.7). Were there to be differences in relative thematic preference displayed by children over adults, we expected to see this for higher ranges or lower ranges exclusively, not for both. We take this as further reason to treat our results cautiously.

While we had expected stronger results, these results do weakly support (and certainly don't oppose) our expectations based on earlier research by Smiley et al. [9] that children exhibit a taxonomic shift between age 6 and the adulthood range that most of our responses came from.

Future attempts at clarifying this should consider either repeating Hebart et al.'s model-creation-and-labelling processing for the responses from college-age students or restricting the questions considered to those with college-age respondents, as it is possible that an expected thematic preference from much-older-adults has contributed to the insignificance of these results. We have not done so for two reasons. First, it would have removed this chapter's connection with the embedding Hebart et al. produced, from which we used the associated validated human dimension labels to generate our taxonomic/thematic labels (see Section 4.4.5.2 and Appendix C). Second, we realized this after we had generated those taxonomic/thematic dimension labels.

Here, we must also mention that we also have implicit assumptions that 1) taxonomic and thematic scores correspond with changes in the object-similarity decision-making thought process and that 2) they capture taxonomic/thematic reasoning at comparable levels.

Addressing the first assumption, on an intuitive level, taxonomic and thematic scores result from taking the maximum of scores correlated with the pre-normalized similarity scores (see Equation 2.6), and as such varying them should impact the model's predictions. Empirically, as well, both scores trend with overall model confusion (see Figure 4.6).

Our second assumption (that our taxonomic and thematic scores respectively capture the full scopes of taxonomic and thematic reasoning at similar levels) is more difficult to check, and we have not done so. A future project may wish to normalize these scores using some external taxonomic or thematic score or proxy. This might include existing scores in the case of taxonomic relations or lexical co-occurrence as a proxy for thematic relations.

4.7 Experiment 2: Age and Absolute Taxonomic and Thematic Relation-Preference Trends

We examine the effect of age on taxonomic- and thematic-relation preferences, where the taxonomic and thematic preferences are measured independently of each other. We do this for different levels of expected taxonomic and thematic strengths.

4.7.1 Experiment Details

For our second experiment, we consider whether children think more taxonomically or thematically than adults, where the taxonomic and thematic preferences are independent of one another. This is in contrast to Experiment 1, which concerned itself with whether children, as compared with adults, exhibited a relatively greater or lesser taxonomic preference than thematic preference.

To capture this, we look at the rates of adult–child response discrepancy at different amounts of taxonomic/thematic confusion. We take the response-difference rates for different amounts of taxonomic and thematic confusion and perform logistic regression to determine overall trends. We use Wald's test to determine significance.

4.7.2 Results

Figure 4.2 and Figure 4.5 give logistic regression plots of the adult-child response-switch rate vs. level of taxonomic or thematic confusion. For the sake of better understanding results, logistic regressions of adult–child response-switch rates vs. overall (total) relative confusion, the largest of factors needing to be controlled for (see Section 4.4.5.5), are provided in Figure 4.3 and Figure 4.4. Regression coefficients and the results of running Wald's test for statistical significance are displayed in Table 4.3 and Table 4.4.

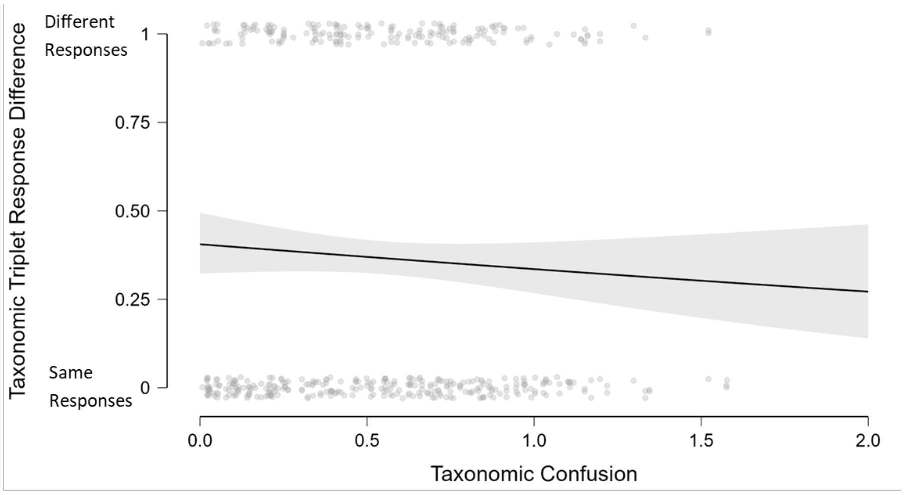


Figure 4.2: A logistic regression of the taxonomic response-difference rate between children and adults on the taxonomically confusing triplets for different levels of taxonomic confusion.

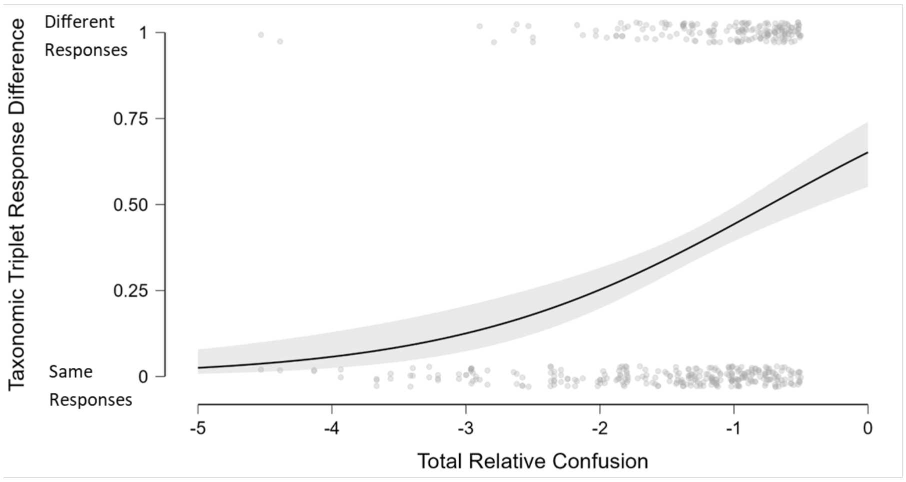


Figure 4.3: A logistic regression of the response-difference rate between children and adults on the taxonomically confusing triplets for different levels of overall relative confusion.

				Wald Test	
	Estimate	StE	z	Wald Statistic	p
Taxonomic Confusion	−0.302	0.282	−1.070	1.145	0.285
Total Relative Confusion	0.856	0.158	5.411	29.279	< .001

Table 4.3: Taxonomic triplet results. Taxonomic confusion versus adult–child response-difference rate logistic regression coefficients and significance test.

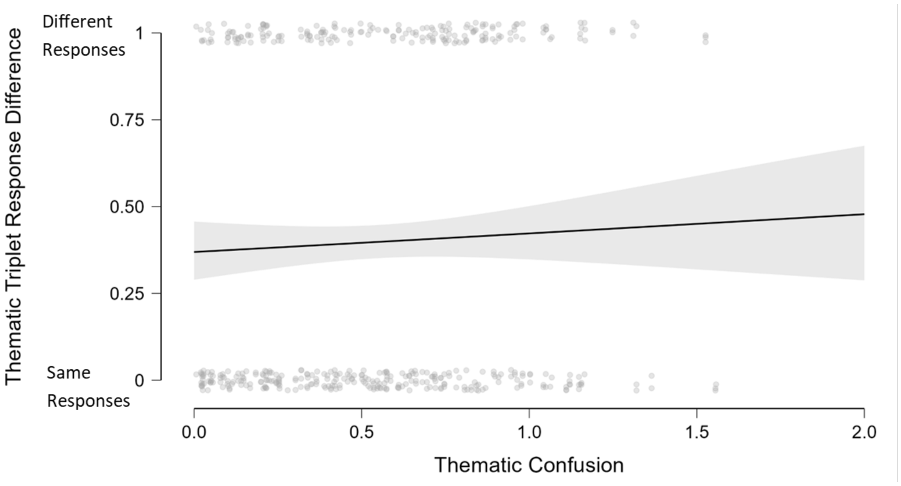


Figure 4.4: A logistic regression of the response-difference rate between children and adults on the thematically confusing triplets for different levels of thematic confusion.

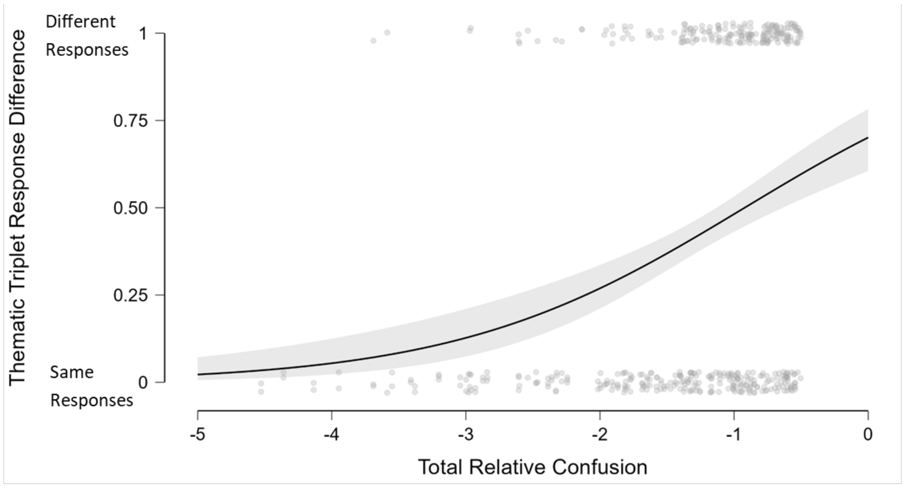


Figure 4.5: A logistic regression of the response-difference rate between children and adults on the thematically confusing triplets for different levels of overall relative confusion.

				Wald Test	
	Estimate	StE	z	Wald Statistic	p
Thematic Confusion	0.224	0.280	0.800	0.639	0.424
Total Relative Confusion	0.926	0.161	5.747	33.033	< .001

Table 4.4: Thematic triplet results. Thematic confusion versus adult–child response-difference rate logistic regression coefficients and significance test.

4.7.3 Discussion

As a recap, in this experiment, we tested for differences between adults (age 18 and up) and children (age 6) in absolute taxonomic and thematic preferences. We did this by choosing triplets where response-switching would be indicative of taxonomic or thematic preferences, the degree to which we referred to as "taxonomic and thematic confusion." Responses for various ranges of taxonomic/thematic confusion, taxonomic/thematic responses-switching, and significant results taxonomic/thematic preferences are displayed in Table 4.2. Across all triplets, we weakly controlled for overall confusion, taxonomic confusion, thematic confusion, thematic affirmation, and taxonomic affirmation; see Section 4.4.6.3.

For our question "how do child preferences change as compared with adults for varying levels of thematic confusion," Experiment 2 ostensibly demonstrates that an increase in the level of thematic confusion results in a greater rate of adult-child response discrepancy, as indicated in Figure 4.5. However, when accounting for overall confusion, this significance disappeared, as seen in Figure 4.4 and Table 4.4. We note here that when we did the controlling for overall confusion, we did so mostly with the intent of controlling for its presence in the pairs for Experiment 1 rather than its presence across different levels of taxonomic/thematic confusion. In any case, the rate of response-switching is explained by the presence of overall confusion. As such, no conclusions can be drawn other than that instances where the model is more likely to disagree with adults are indicative that children might disagree with those same adults.

Even were we to have found significant results, however, the results for the question "how do child preferences change as compared with adults for varying levels of taxonomic confusion" give us reason to pause. Again, here the results are entirely explained by total relative confusion (see Figure 4.3 and Table 4.3), but the observed adult-child response-discrepancy rate *decreases* with increasing taxonomic confusion (see Figure 4.2). This is unexpected regardless of how valid a measure of taxonomic preference our taxonomic scores are, as any confusion scores should still correlate with increasing response switches. While simple checks for correlation with the overall relative confusion (Figure 4.6) did not yield any insight, we ultimately still suspect this is due to some form of correlation with the controls. Considering overall relative confusion's predictive power, future work should still make constraining it across triplets of different taxonomic/thematic confusion a priority.

One other untouched-upon note is that we are making an implicit assumption that our measures of taxonomic and thematic reasoning correlate with adjustments in object-similarity determination and that both capture taxonomic and thematic reasoning at similar levels. We have good evidence for the former, but not much for the latter; see the discussion at the end of Section 4.6.3 for details and possible future directions.

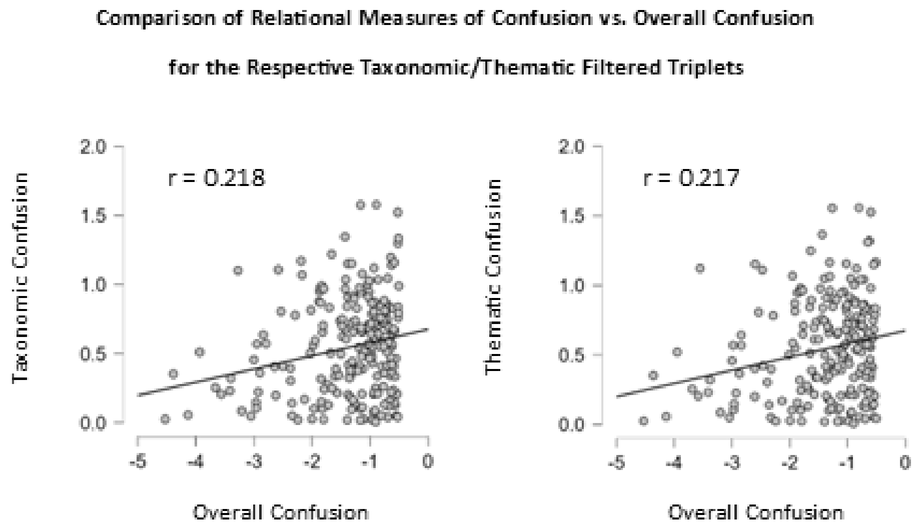


Figure 4.6: A logistic regression of the response-switch rate between children and adults for different levels of overall relative confusion. Each point is a triplet, and the x-axis gives its overall confusion score, while the y-axis gives either its taxonomic or thematic confusion. Notice the similar trendlines. Here, "taxonomic confusion" and "thematic confusion" refer to our taxonomic and confusion scores. (Section 4.4.5.3 for taxonomic/thematic confusion and Section 4.4.5.5 for overall confusion details).

Chapter 5

Conclusion

This chapter concludes the thesis. Here, we reiterate our driving questions and experiment procedures, state our key findings, clarify their limitations, and offer suggestions for future work. We end by summarizing and offering final thoughts about the work presented in this thesis.

5.1 Thesis Questions

Our driving thesis questions are as follows:

1. When considering a set of human-interpretable, computationally-derived similarity dimensions, do adults aged 25–35 and 50–60 prioritize those human-interpretable similarity dimensions differently when comparing objects? In what ways?

2. Do children (age 6) exhibit stronger taxonomic- or thematic-relation preferences than adults, where one: the taxonomic and thematic preferences are measured relative to each other, and two: the taxonomic and thematic preferences are measured independently?

5.2 Chapter 3
Adult Age and Type-of-Similarity Preference

5.2.1 Motivation

In Chapter 3, we examined our first question, which concerned adult ages and object-similarity measure preferences. Our motivations for addressing this question were threefold.

First, a scarcity of prior results: prior research has generally focused on the relationship between adult age and broad classes of object-similarity measures, rather than on more specific ones, such as "metallic." Second, novelty of technique: to our knowledge, very little work, if any, has connected similarity-task-derived computational measures to differently aged adults' preferences for those measures. Third, expanding available resources: determining age-based preferences for these similarity measures contributes to the THINGS initiative's [15] shared body of knowledge.

5.2.2 Approach

We obtained measures of specific types of object-similarity prioritization, or ***likenesses***, for 25–35-year-olds and 50–60-year-olds. We did this by modifying the object-comparison-performing model of Hebart et al. [13], the dimensions of which correspond with object-comparison relations and have human-identifiable labels. We refer to these dimensions as ***likeness dimensions***, with the associated type of relation being the ***dimensional likeness***. Our modified model used a layer of ***likeness-preference weights*** that rescaled the prominence of the relations in the original model when determining object similarities. We learned these preference weights for each age group. Finally, we performed hypothesis testing to see if the preference weights differed between groups beyond random chance.

5.2.3 Results

We found many statistically significant differences between younger and older adults concerning their type-of-similarity (likeness) preferences in performing the odd-one-out task. Specifically, we found significant differences in the likeness-preference weights for 23 of the 49 likeness dimensions. These are presented in tabular form in Table 3.1 and graphically in Appendix A. Most notably, we found indications that the three largest results among those were where older adults exhibited a preference for perceptual features over younger adults, namely for "colorful," "disc-shaped/round," and "long/thin."

The other largest age-group-discrepant result was a relatively greater preference for "medicine-related" among 25–35-year-olds. We interpreted this as possibly indicating that older individuals may have a more nuanced, or at least different, interpretation of things being medicine-related than the population as a whole.

For a more careful treatment of these conclusions, as well as the complete set of results, see Section 3.5.2.

5.2.4 Limitations and Future Work

The results of Chapter 3 are subject to one major limitation: while differences in preference for an object-similarity (likeness) dimension do reflect a given dimension's lesser importance for that group, this can be due to one of two reasons. First, the most useful reason: a lesser/greater usage of a likeness dimension could reflect that group's lesser/greater usage of the corresponding natural dimensional likeness. The second interpretation, however, is more limited: a lesser/greater usage of a likeness dimension could indicate that the group uses the corresponding dimensional likeness in a different way than it was encoded, i.e., they may use some of the more fine-grained aspects of similarities that comprise the likeness dimension, but not others.

Future psychology work should interrogate the result indicating that older adults exhibit preferences for the perceptual “colorful,” “disc-shaped/round,” and “long/thin” object-similarity features. Future computing science work may use our setup to examine object-similarity preferences between other pairs of groups in the THINGS response dataset with sufficient data, such as “male” and “female” groups.

5.3 Chapter 4 Children vs. Adults in Taxonomic and Thematic Prioritization

5.3.1 Motivation

In Chapter 4, we examined our second question, which concerned discrepancies between children and adults on taxonomic- and thematic-relation preferences. Our primary motivations for doing so were twofold.

First, we sought to corroborate existing results from studies where humans manually selected questions for exploring this issue by instead using automatically-generated questions from a computational model. Prior research [9] found that children exhibit a shift from thematic preference to taxonomic preference between ages 6 and 10 and that this change in preference remains into part of adulthood. Second, we sought to derive measures of age-based changes in taxonomic and thematic preferences independent of one another. Prior research has instead largely focused on the two relative to one other [8].

5.3.2 Approaches

In Chapter 4, we took two approaches: one to corroborate existing results concerning relative changes in taxonomic and thematic preferences, and a second to determine absolute changes in taxonomic and thematic preferences.

Both approaches began as follows. To start, we had a set of adult responses to an odd-one-out-among-three object-comparison task. Each response was taken on a set of three objects, or object triplet, that defined an odd-one-out question at hand. We used a computational model to give these questions ***taxonomic and thematic confusion*** scores. These scores indicate the extent that the deviation of someone's response from the original adult data would signal a preference for taxonomic or thematic relations relative to those adult respondents.

For our first experiment, to corroborate existing results on the connection between age and changes in relative taxonomic and thematic preferences, we generated pairs of triplets such that one member of each pair had a level of taxonomic confusion similar to the other member's level of thematic confusion. At different levels of taxonomic and thematic confusion, we then recorded whether children had different responses than adults. A higher number of child–adult response discrepancies for taxonomically confusing triplets than thematically confusing triplets indicated a relative taxonomic-over-thematic preference (and vice versa).

For our second experiment, to determine changes in absolute terms ("absolute" being taxonomic and thematic preference changes independent of one another; both were still relative to the adult respondents), we selected triplets with varying levels of taxonomic and thematic confusion and performed regression analysis. To minimize the number of participants needed, we combined this with the previous experiment's procedure to get samples for determining both relative and absolute trends simultaneously.

5.3.3 Results

Our first experiment aimed to corroborate existing results about relative changes in taxonomic and thematic preferences with age. These results are presented in Table 4.2 and Appendix D.

Based on prior research, we expected a thematic-to-taxonomic relative preference shift. We did find some support for this; however, the support was quite limited. Most concerningly, we found support for this for large and small levels of taxonomic/thematic confusion but not for medium levels, which we find to be counterintuitive. A complete discussion of this can be found in Section 4.6.3.

For our second experiment, we strove to find absolute trends in taxonomic and thematic preferences, rather than trends relative to one another. The results are presented in Section 4.7.2.

We insufficiently controlled for a factor called ***overall confusion*** (Section 4.4.5.5), which correlated with our results to the point that our work does not answer our guiding research question. Instead, we showed that when overall confusion increased for a triplet-adult response pair (that is, the model of Hebart et al. has increasing reason to disagree with the adult response), a child was also more likely to disagree with the adult response, which was to be expected. This situation is a consequence of us combining our sampling for both questions; when doing so, we prioritized the relative taxonomic/thematic preference setup in controlling for various factors for our choice of questions to administer. A complete discussion of this is found in Section 4.7.3.

5.3.4 Limitations and Future Work

Chapter 4 has heavily limited results.

For the relative-preference trends, we suspect the limited results to be due to two factors. First, our choices of taxonomic and thematic confusion involved a number of not-thoroughly-tested ad-hoc decisions. Verifying the validity of and improving upon these choices to produce better numeric measures of taxonomic and thematic similarities are potential future projects. Second, the inclusion of older adults may have influenced the taxonomic and thematic trends, as significantly older adults may exhibit more thematic thinking than younger adults. Those working on a future project may wish to either determine a way to control for these or abandon the original Hebart et al. model and train a new one, although the latter decision would make the project offer less utility with regards to the THINGS initiative.

The absolute preference results are subject to the same concerns regarding the (necessary given the time-frame for this thesis) somewhat makeshift choices in determining taxonomic and thematic scores. Additionally, however, these results suffered from our decision to collect data for both the relative and absolute trends simultaneously, resulting in weaker controlling for the factor called "overall confusion" in the process. Future work should avoid combining the absolute-preference experiment with the relative-preference-results experiment, controlling for overall confusion (and the other control factors) in triplet selection for each in separate setups.

5.4 Closing Summary and Final Thoughts

5.4.1 Closing Summary

In this thesis, we used a computational object-similarity model to find differences between age groups in object-similarity judgment. We specifically explored this in two ways. First, we looked at the differences between adults aged 25–35 and adults aged 50–60 in their usage of fine-grained aspects of similarity. Second, we examined the differences between children (age 6) and adults in their preferences for taxonomic and thematic relations, both with the relation strengths measured relative to each other and individually.

Concerning the older and younger adults and fine-grained object-similarity features, we found significant novel differences between each age group's preferences for 23 object-similarity features. This adds to the body of literature on age-based differences in object-similarity judgment, and our observed trend of older adults having much higher preference scores than younger adults in several perceptual dimensions warrants future exploration.

Concerning adults' and children's taxonomic and thematic preferences, we obtained weakly corroborative results on existing connections between age and relative taxonomic/thematic preferences. We did this with novel computational-model-based methods. We also described future steps in determining absolute taxonomic/thematic preferences and a procedure for obtaining them.

Finally, in both projects, we expanded on the THINGS initiative with new results. In particular, we determined how differently each dimension of the THINGS object-similarity model explains the object-similarity judgments of 25–35-year-olds and of 50–60-year-olds, with methodology adaptable to other groups in the future.

5.4.2 Final Thoughts

Interpretable model embeddings like Hebart et al.'s offer useful proxies for evaluating thought processes. Going forward, I am keen to see what comes of the observed perceptual adult-age object-similarity-preference differences, particularly how they might be elaborated on through results using other techniques. Although our second project's results were inconclusive, the relation-scoring portion deserves additional focus, after which more concrete results may be obtainable. While it can be challenging to precisely determine the manners by which people compare objects, my hope is that this thesis contributes to our understanding of them.

Bibliography

[1] G. Desmarais, M. C. Pensa, M. J. Dixon, and E. A. Roy, "The importance of object similarity in the production and identification of actions associated with objects," en, *Journal of the International Neuropsychological Society*, vol. 13, no. 6, pp. 1021–1034, Nov. 2007, ISSN: 1355-6177, 1469-7661. DOI: 10.1017/S1355617707071287. [Online]. Available: https://www.cambridge.org/core/product/identifier/S1355617707071287/type/journal_article.

[2] E. M. Markman and J. E. Hutchinson, "Children's sensitivity to constraints on word meaning: Taxonomic versus thematic relations," en, *Cognitive Psychology*, vol. 16, no. 1, pp. 1–27, Jan. 1984, ISSN: 00100285. DOI: 10.1016/0010-0285(84)90002-1. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0010028584900021 (visited on 04/21/2022).

[3] L. W. Barsalou, "Ad hoc categories," en, *Memory & Cognition*, vol. 11, no. 3, pp. 211–227, May 1983, ISSN: 0090-502X, 1532-5946. DOI: 10.3758/BF03196968. [Online]. Available: http://link.springer.com/10.3758/BF03196968 (visited on 08/20/2022).

[4] P. Tirilly, X. Mu, C. Huang, I. Xie, W. Jeong, and J. Zhang, "On the consistency and features of image similarity," en, in *Proceedings of the 4th Information Interaction in Context Symposium on - IIIX '12*, Nijmegen, The Netherlands: ACM Press, 2012, pp. 164–173, ISBN: 978-1-4503-1282-0. DOI: 10.1145/2362724.2362754. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2362724.2362754 (visited on 08/10/2022).

[5] Z. Lun, E. Kalogerakis, and A. Sheffer, "Elements of style: Learning perceptual shape style similarity," en, *ACM Transactions on Graphics*, vol. 34, no. 4, pp. 1–14, Jul. 2015, ISSN: 0730-0301, 1557-7368. DOI: 10.1145/2766929. [Online]. Available: https://dl.acm.org/doi/10.1145/2766929 (visited on 08/23/2022).

[6] L. J. Speed, J. Chen, F. Huettig, and A. Majid, "Do classifier categories affect or reflect object concepts?" In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*, Philadelphia, PA: Cognitive Science Society, 2016, pp. 2267–2272.

[7] J. B. Proffitt, J. D. Coley, and D. L. Medin, "Expertise and category-based induction," en, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 26, no. 4, pp. 811–828, 2000, ISSN: 1939-1285, 0278-7393. DOI: 10.1037/0278-7393.26.4.811. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/0278-7393.26.4.811 (visited on 08/21/2022).

[8] D. Mirman, J.-F. Landrigan, and A. E. Britt, “Taxonomic and thematic semantic systems,” en, *Psychological Bulletin*, vol. 143, no. 5, pp. 499–520, May 2017, ISSN: 1939-1455, 0033-2909. DOI: 10.1037/bul0000092. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/bul0000092 (visited on 04/20/2022).

[9] S. S. Smiley and A. L. Brown, “Conceptual preference for thematic or taxonomic relations: A nonmonotonic age trend from preschool to old age,” en, *Journal of Experimental Child Psychology*, vol. 28, no. 2, pp. 249–257, Oct. 1979, ISSN: 00220965. DOI: 10.1016/0022-0965(79)90087-0. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0022096579900870 (visited on 04/21/2022).

[10] C. Berger and S. Donnadieu, “Visual/auditory processing and categorization preferences in 5-year-old children and adults,” en, *Current psychology letters*, no. Vol. 24, Issue 2, 2008, Sep. 2008, ISSN: 1376-2095, 1379-6100. DOI: 10.4000/cpl.3673. [Online]. Available: http://journals.openedition.org/cpl/3673 (visited on 04/20/2022).

[11] U. Hahn, “Similarity,” en, *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 5, no. 3, pp. 271–280, May 2014, ISSN: 19395078. DOI: 10.1002/wcs.1282. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/wcs.1282 (visited on 07/27/2022).

[12] A. Tversky, “Features of similarity.,” *Psychological review*, vol. 84, no. 4, p. 327, 1977.

[13] M. N. Hebart, C. Y. Zheng, F. Pereira, and C. I. Baker, “Revealing the multidimensional mental representations of natural objects underlying human similarity judgements,” en, *Nature Human Behaviour*, vol. 4, no. 11, pp. 1173–1185, Nov. 2020, ISSN: 2397-3374. DOI: 10.1038/s41562-020-00951-3. [Online]. Available: https://www.nature.com/articles/s41562-020-00951-3 (visited on 07/26/2022).

[14] M. N. Hebart *et al.*, “THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images,” en, *PLOS ONE*, vol. 14, no. 10, F. A. Soto, Ed., e0223792, Oct. 2019, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0223792. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0223792 (visited on 07/26/2022).

[15] *The THINGS Initiative*, https://web.archive.org/web/20220824022100/https://things-initiative.org/, Accessed: 202-08-23.

[16] D. Gentner and A. B. Markman, “Structure mapping in analogy and similarity.,” *American psychologist*, vol. 52, no. 1, p. 45, 1997.

[17] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem, “Basic objects in natural categories,” en, *Cognitive Psychology*, vol. 8, no. 3, pp. 382–439, Jul. 1976, ISSN: 00100285. DOI: 10.1016/0010-0285(76)90013-X. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/001002857690013X (visited on 04/20/2022).

[18] P. University, *WordNet Online*, https://wordnet.princeton.edu/.

[19] T. Mikolov, K. Chen, G. Corrado, and J. Dean, *Efficient estimation of word representations in vector space*, 2013. DOI: 10.48550/ARXIV.1301.3781. [Online]. Available: https://arxiv.org/abs/1301.3781.

[20] S. Kalénine, D. Mirman, E. L. Middleton, and L. J. Buxbaum, "Temporal dynamics of activation of thematic and functional knowledge during conceptual processing of manipulable artifacts.," en, *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 38, no. 5, pp. 1274–1295, 2012, ISSN: 1939-1285, 0278-7393. DOI: 10.1037/a0027626. [Online]. Available: http://doi.apa.org/getdoi.cfm?doi=10.1037/a0027626 (visited on 07/05/2022).

[21] P. J. Bauer and J. M. Mandler, "Taxonomies and triads: Conceptual organization in one- to two-year-olds," en, *Cognitive Psychology*, vol. 21, no. 2, pp. 156–184, Apr. 1989, ISSN: 00100285. DOI: 10.1016/0010-0285(89)90006-6. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0010028589900066 (visited on 04/20/2022).

[22] L. Fenson, D. Vella, and M. Kennedy, "Children's Knowledge of Thematic and Taxonomic Relations at Two Years of Age," en, *Child Development*, vol. 60, no. 4, p. 911, Aug. 1989, ISSN: 00093920. DOI: 10.2307/1131032. [Online]. Available: https://www.jstor.org/stable/1131032?origin=crossref (visited on 04/21/2022).

[23] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics*, vol. 5, no. 4, pp. 115–133, Dec. 1943, ISSN: 1522-9602. DOI: 10.1007/BF02478259. [Online]. Available: https://doi.org/10.1007/BF02478259.

[24] G. E. Alefeld, F. A. Potra, and Y. Shi, "Algorithm 748: Enclosing zeros of continuous functions," en, *ACM Transactions on Mathematical Software*, vol. 21, no. 3, pp. 327–344, Sep. 1995, ISSN: 0098-3500, 1557-7295. DOI: 10.1145/210089.210111. [Online]. Available: https://dl.acm.org/doi/10.1145/210089.210111 (visited on 08/24/2022).

[25] R. R. Wilcox, *Introduction to robust estimation and hypothesis testing*, en, 4th edition. Waltham, MA: Elsevier, 2016, ISBN: 978-0-12-804733-0.

[26] V. Kuperman, H. Stadthagen-Gonzalez, and M. Brysbaert, "Age-of-acquisition ratings for 30,000 English words," en, *Behavior Research Methods*, vol. 44, no. 4, pp. 978–990, Dec. 2012, ISSN: 1554-3528. DOI: 10.3758/s13428-012-0210-4. [Online]. Available: http://link.springer.com/10.3758/s13428-012-0210-4 (visited on 07/26/2022).

[27] L. M. Stoinski, J. Perkuhn, and M. N. Hebart, "THINGS+: New Norms and Metadata for the THINGS Database of 1,854 Object Concepts and 26,107 Natural Object Images," PsyArXiv, preprint, Jul. 2022. DOI: 10.31234/osf.io/exu9f. [Online]. Available: https://osf.io/exu9f (visited on 08/25/2022).

[28] *Wiktionary*, https://en.wiktionary.org/.

[29] *Wiktionary.ru*, https://ru.wiktionary.org/.

[30] K. Lange, S. Kühn, and E. Filevich, ""Just Another Tool for Online Studies" (JATOS): An Easy Solution for Setup and Management of Web Servers Supporting Online Studies," en, *PLOS ONE*, vol. 10, no. 6, D. Margulies, Ed., e0130834, Jun. 2015, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0130834. [On-

line]. Available: https://dx.plos.org/10.1371/journal.pone.0130834 (visited on 08/25/2022).

Appendix A: Optimal Relation-Preference Weight Sampling and Statistics for the 49 Dimensional Likenesses

This appendix has the results for the experiments of Chapter 3. The procedures for these are detailed in Section 3.4.8 and the results are discussed in Section 3.5.2.

Section A.1 contains a pair of graphs per likeness dimension. The upper graphs among these contain the main statistical test results (Section 3.4.8.1) and confidence intervals that contrast randomized preference differences with the age-group preference difference observed on the full sets of both age groups' odd-one-out responses. The lower graphs contain confidence intervals (Section 3.4.8.2) that additionally incorporate bootstrapped differences between subsamples of the full sets of both age groups' odd-one-out responses.

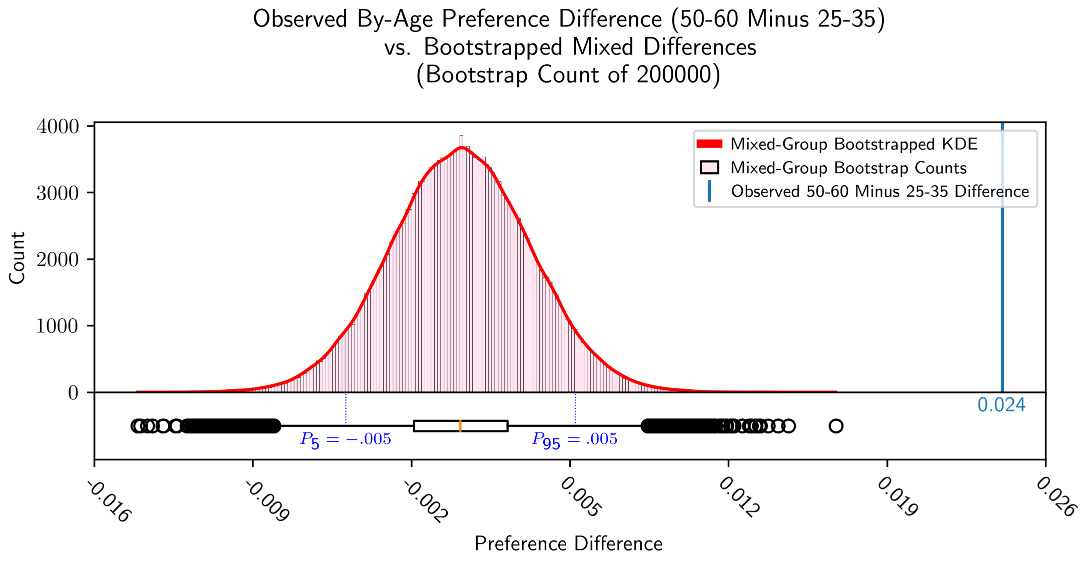
Section A.2 contains the distributions of bootstrapped likeness-preference weights for each age group.

A.1 Age- and Random-Group Preference Difference Bootstrapping and Tests

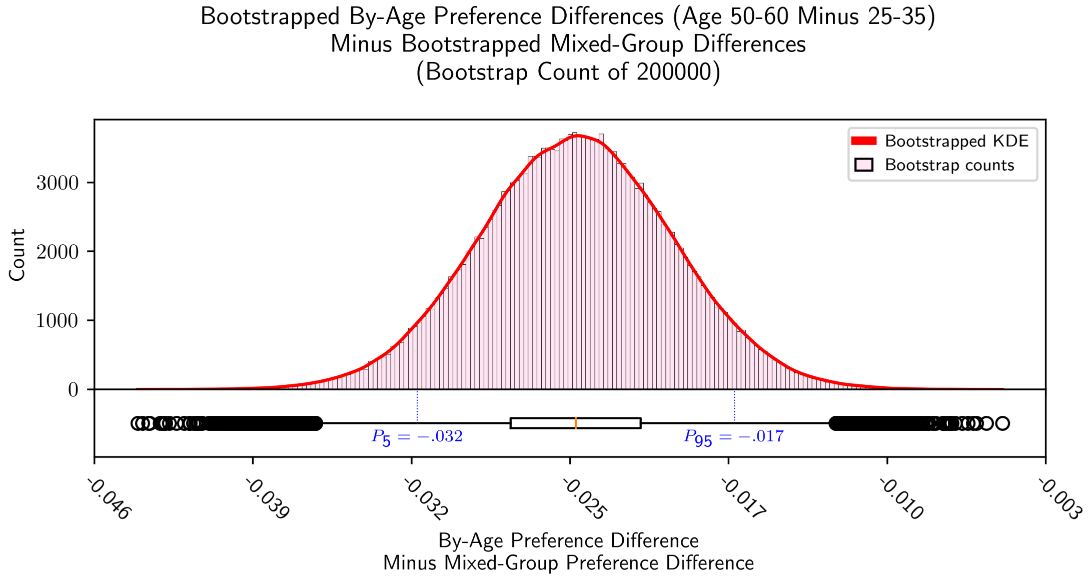
The upper graphs contain Chapter 3's major statistical significance findings. A given upper graph tests the probability of observing the full-response-set age-group preference difference (blue line) under the hypothesis that it was sampled from the distribution of randomized groups' preference differences (pink/red distribution). The x-axis gives the difference in preference weights between each age group. Consequently, the blue line lying right of zero indicates a relatively greater preference by 50–60-year-olds for the given dimension than by 25–35-year-olds, while the blue line lying to the left would denote the opposite. The blue line lying within the pink distribution indicates that the observed all-response preference difference between age groups was not significant. Further details of this are found in Section 3.4.8.1.

A given bottom graph contains bootstrapped differences between the preferences of each age group (50–60 minus 25–35) minus random bootstrapped differences. As such, this distribution lying right of zero indicates a relatively greater preference by 50–60-year-olds for the dimension not explained by chance, whereas this distribution lying left of zero indicates a relatively greater preference by 25—35-year-olds. The more this distribution overlaps zero, the more it indicates a lack of significant difference between the age groups. The confidence interval for this overlapping zero is similar to, but not quite the same as, the difference between the two age groups' full-response-set preferences being insignificant under the upper graph's statistical test; see Section 3.4.8.2 for details.

Dimension 1: made of metal/artificial/hard


Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)

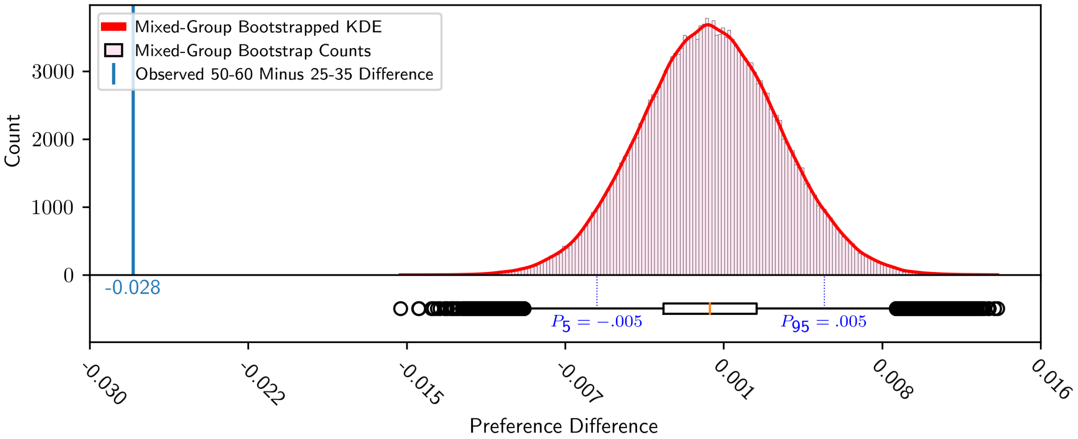
$p = 0.0$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution


Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)

$\text{CI}_{90} = [-0.032, -0.017]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

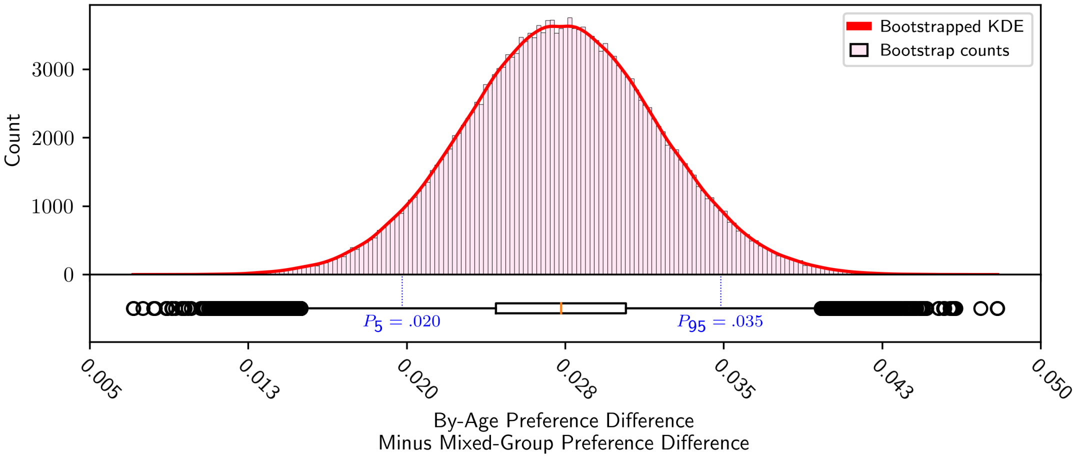
Dimension 2: food-related/eating-related/kitchen-related

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



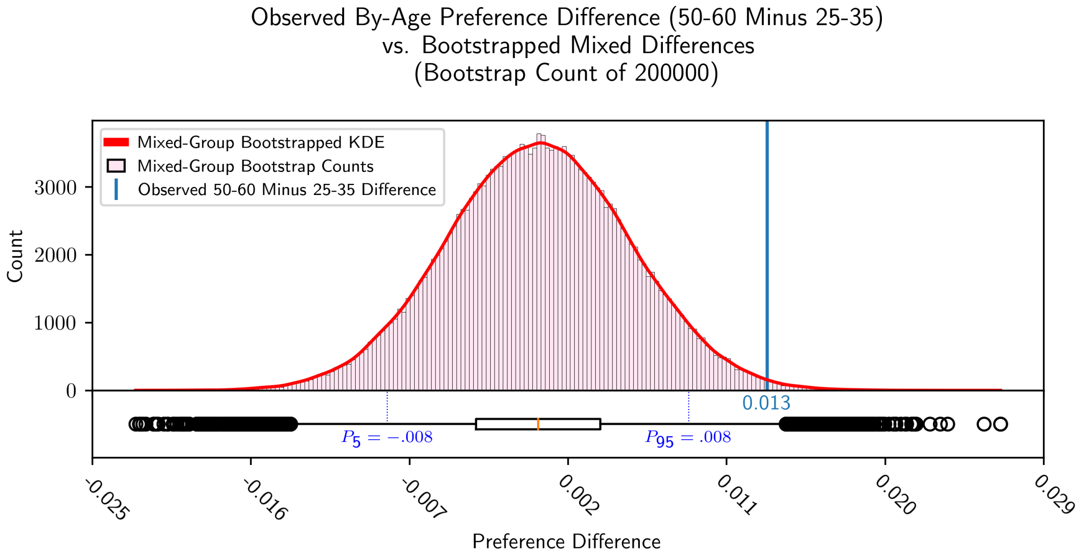
$p = 0.0$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)

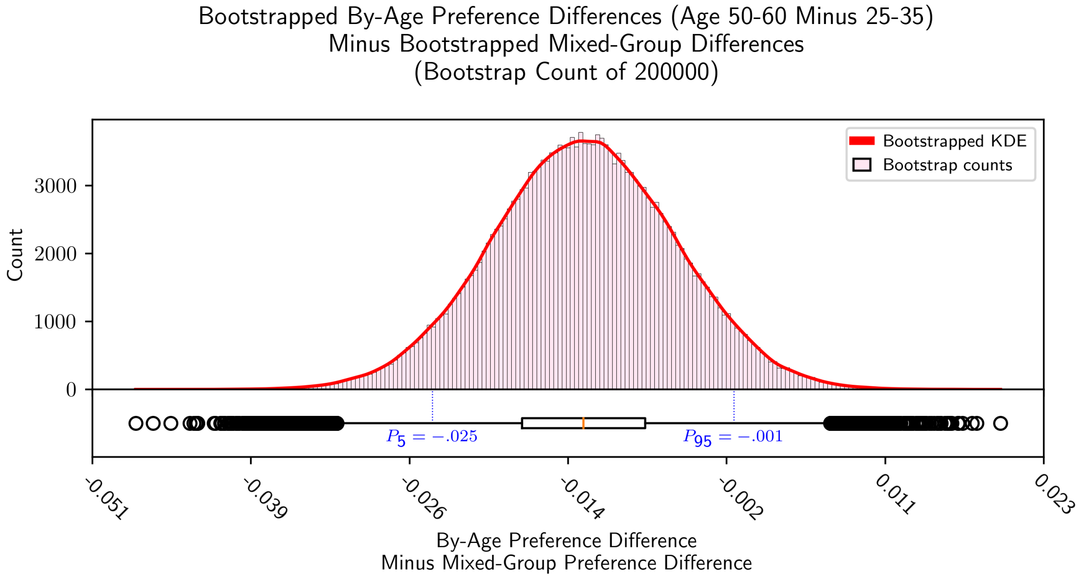


$\mathrm{CI}_{90} = [0.020, 0.035]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

Dimension 3: animal-related/organic

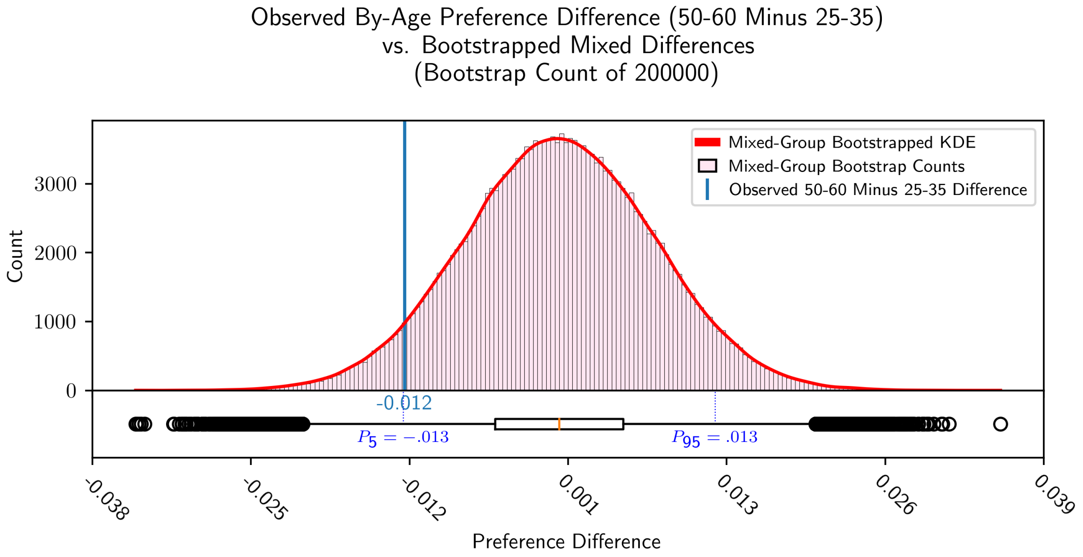


$p = 0.01177$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

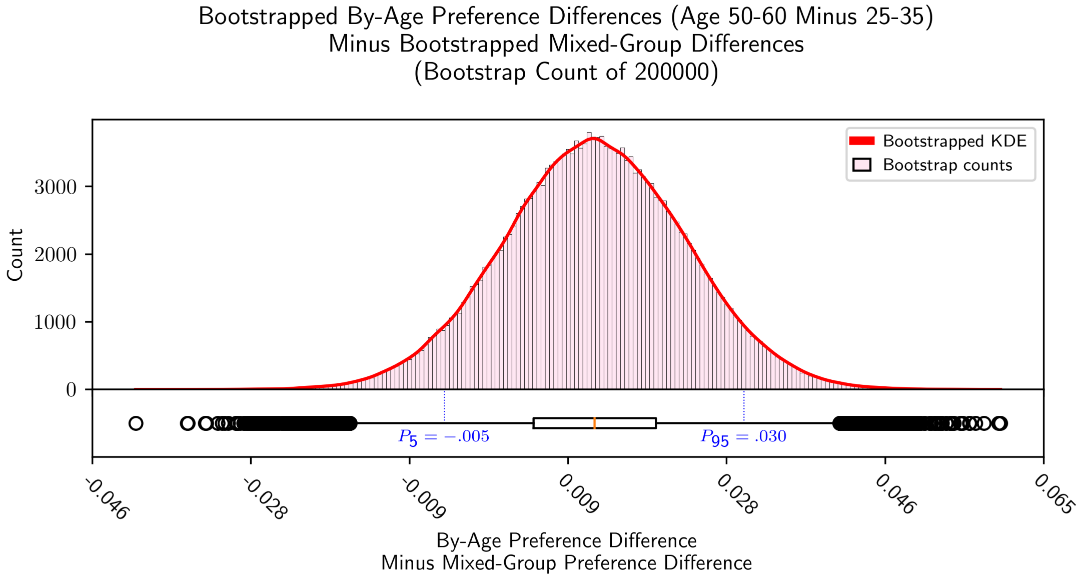


$\mathrm{CI}_{90} = [-0.025, -0.001]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

Dimension 4: clothing-related/fabric/covering



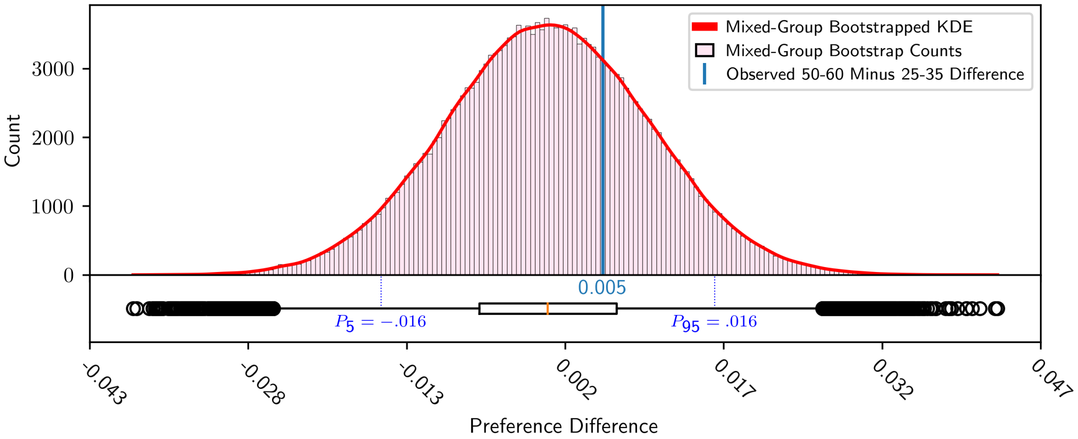
$p = 0.10233$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution



$\mathrm{CI}_{90} = [-0.005, 0.030]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

Dimension 5: furniture-related/household-related/artifact

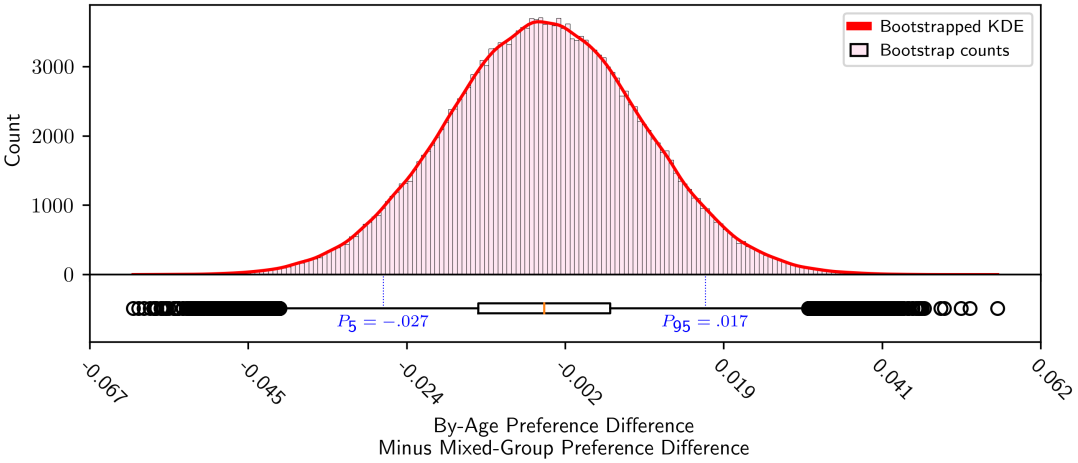
Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.587125$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

---

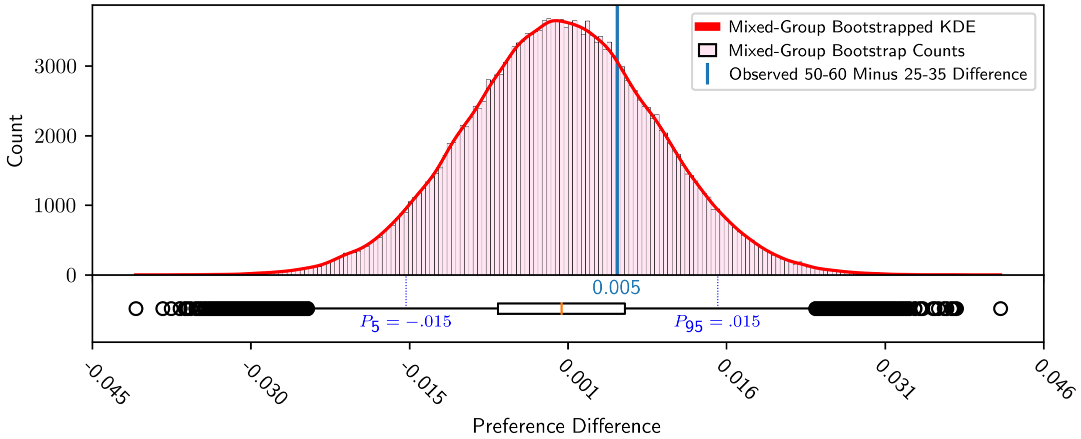
Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



$\mathrm{CI}_{90} = [-0.027, 0.017]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

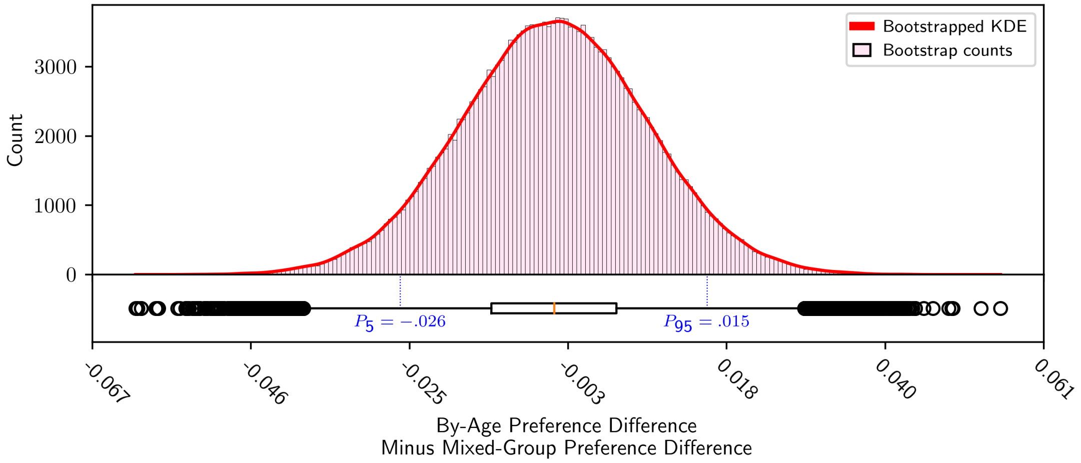
Dimension 6: plant-related/green

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.55342$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

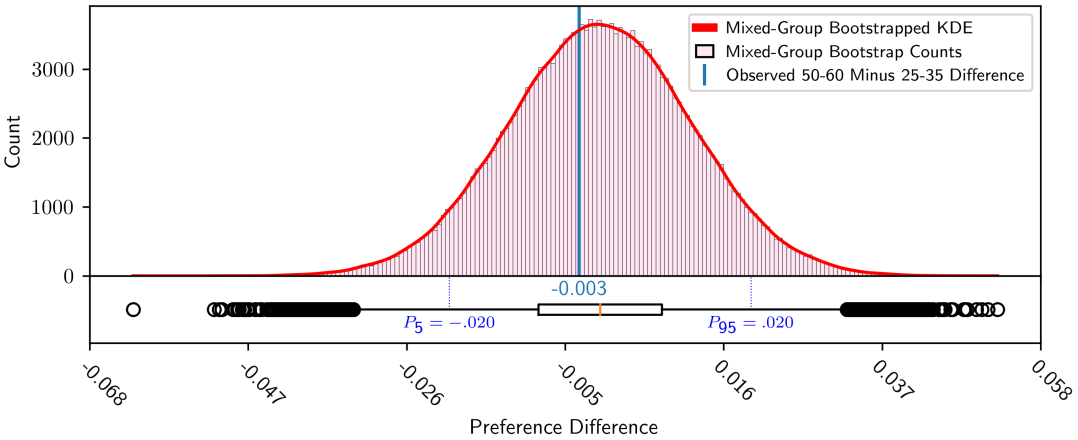
Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



$\mathrm{CI}_{90} = [-0.026, 0.015]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

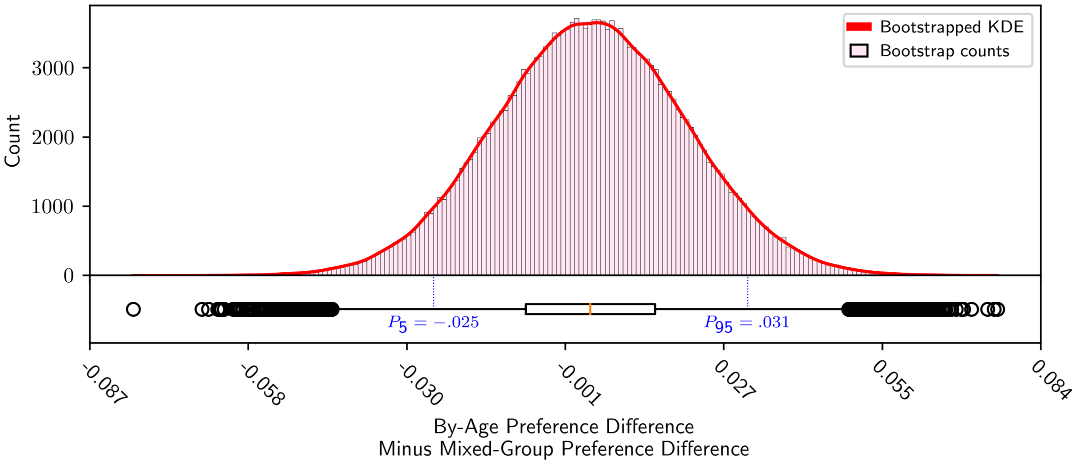
Dimension 7: outdoors-related

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.823085$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

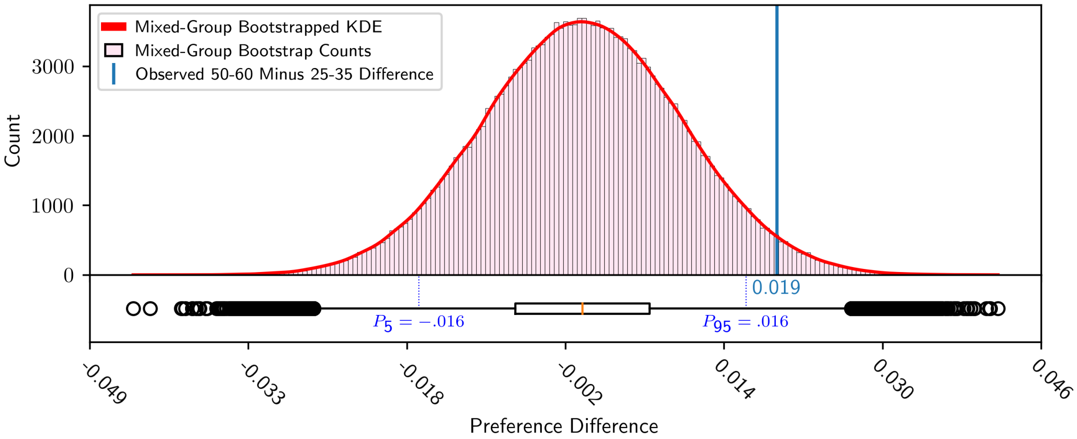
Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



$\mathrm{CI}_{90} = [-0.025, 0.031]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

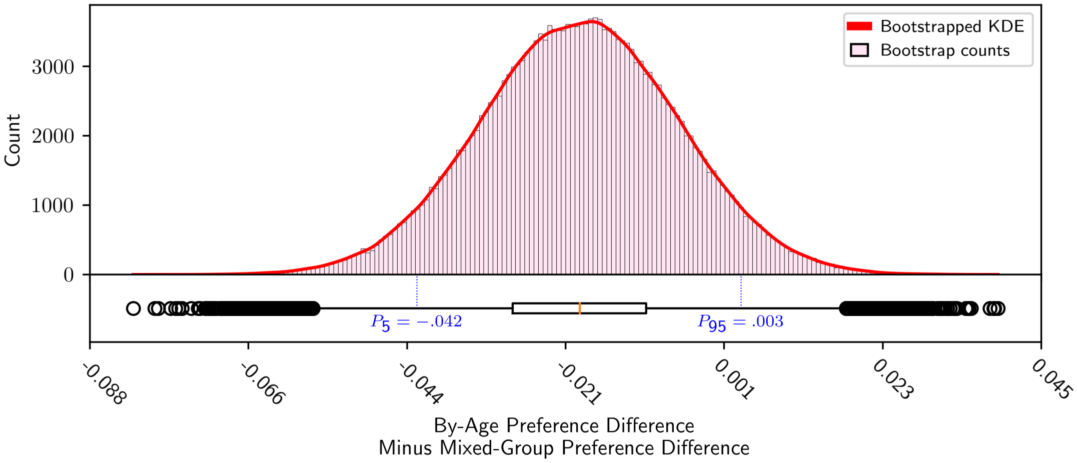
Dimension 8: transportation/motorized/dynamic

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.04979$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

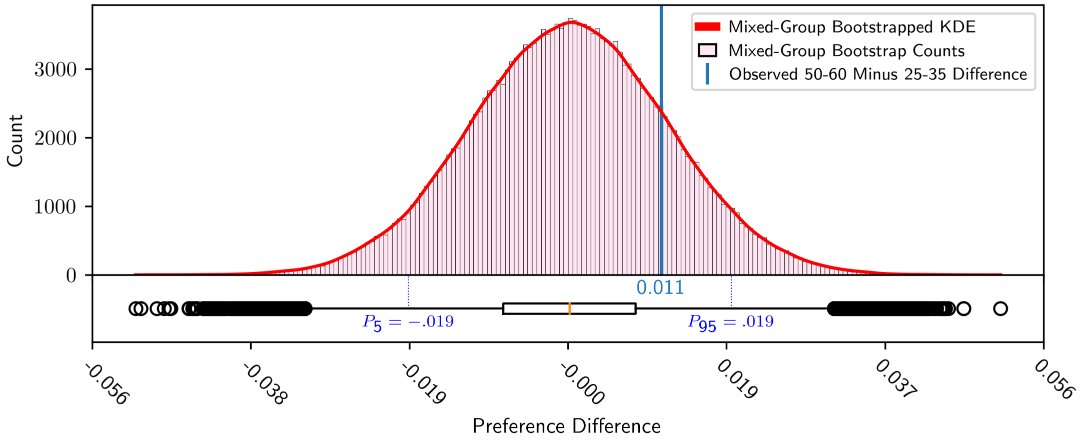
Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



$\mathrm{CI}_{90} = [-0.042, 0.003]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

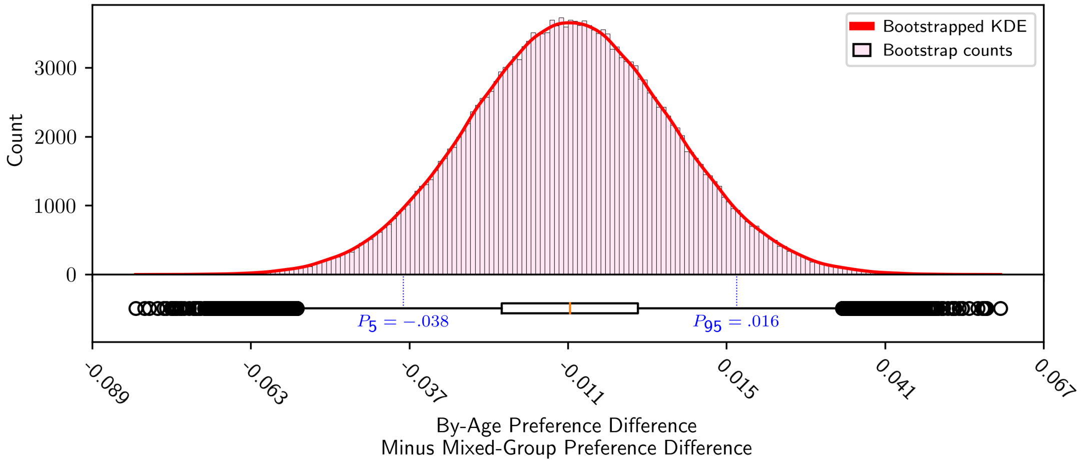
Dimension 9: wood-related/brownish

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.349355$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

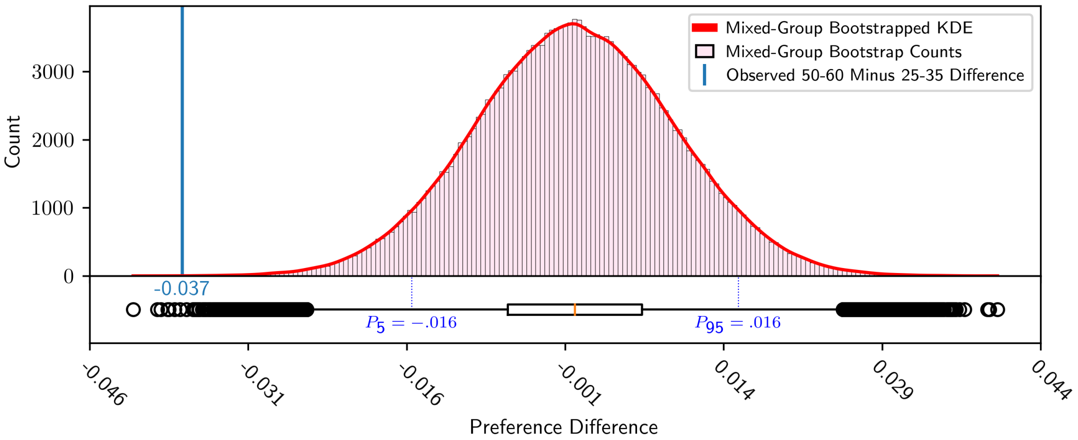
Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



$\mathrm{CI}_{90} = [-0.038, 0.016]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

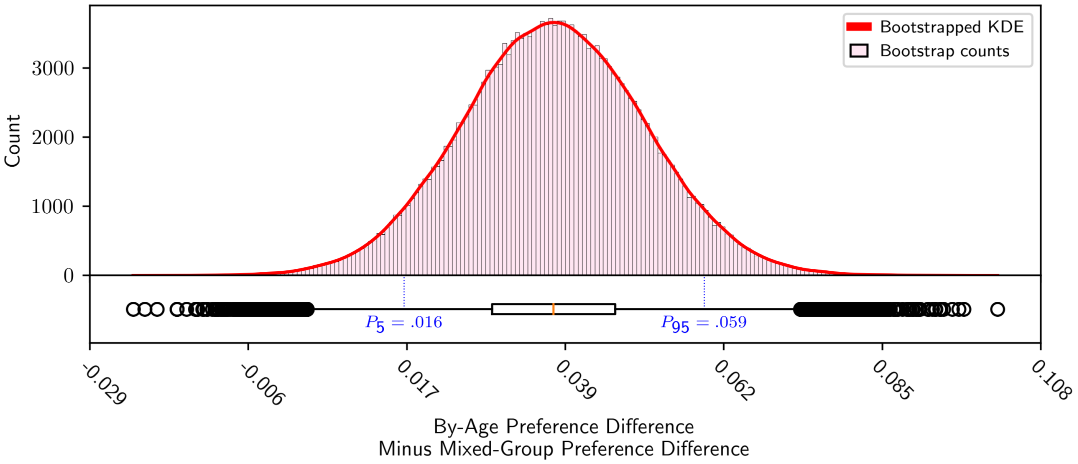
Dimension 10: body part-related

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 5e - 05$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

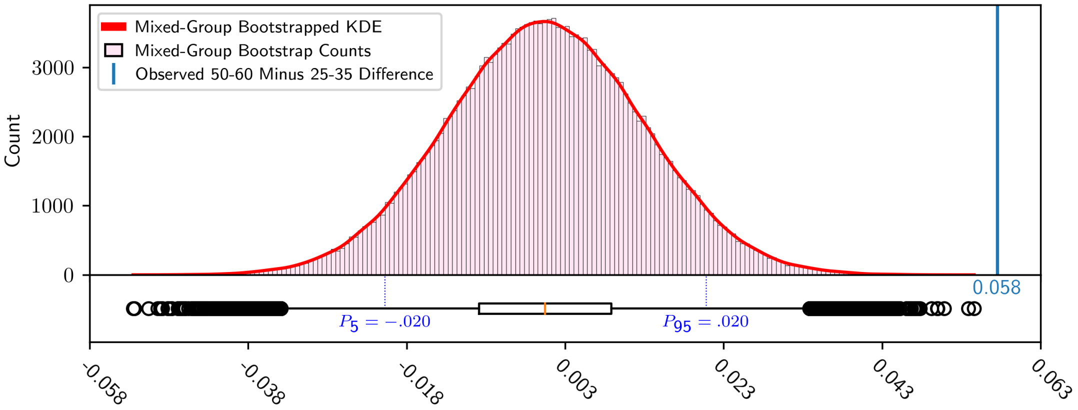
Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



$\mathrm{CI}_{90} = [0.016, 0.059]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

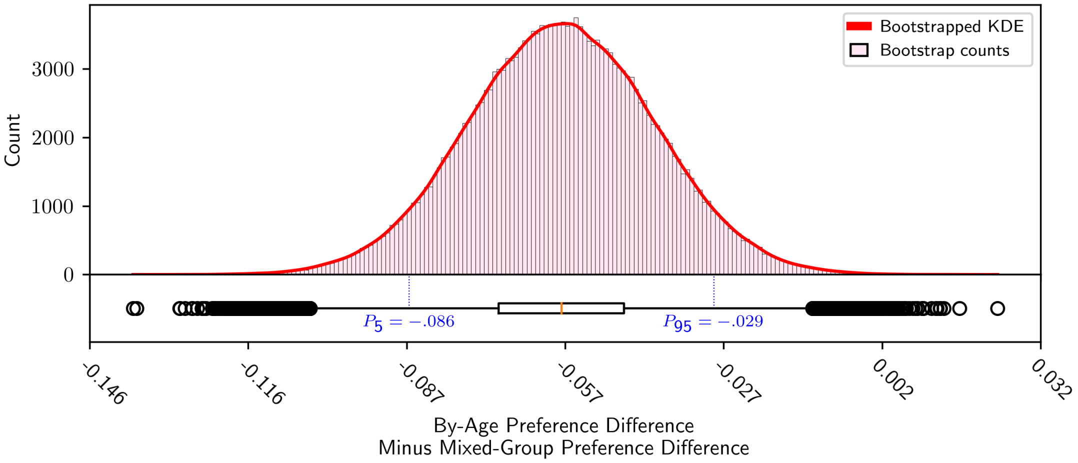
Dimension 11: colorful

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



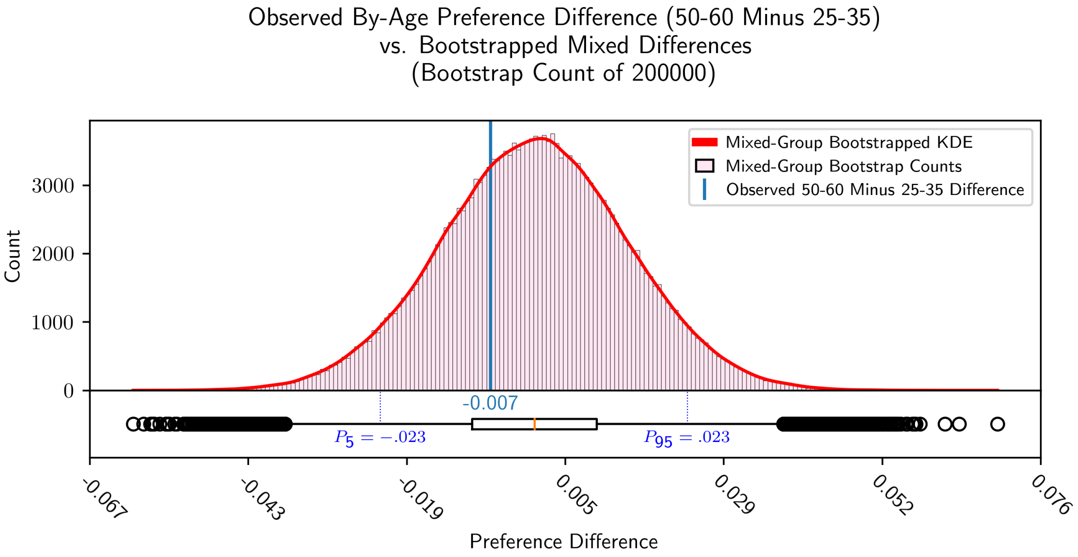
$p = 0.0$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



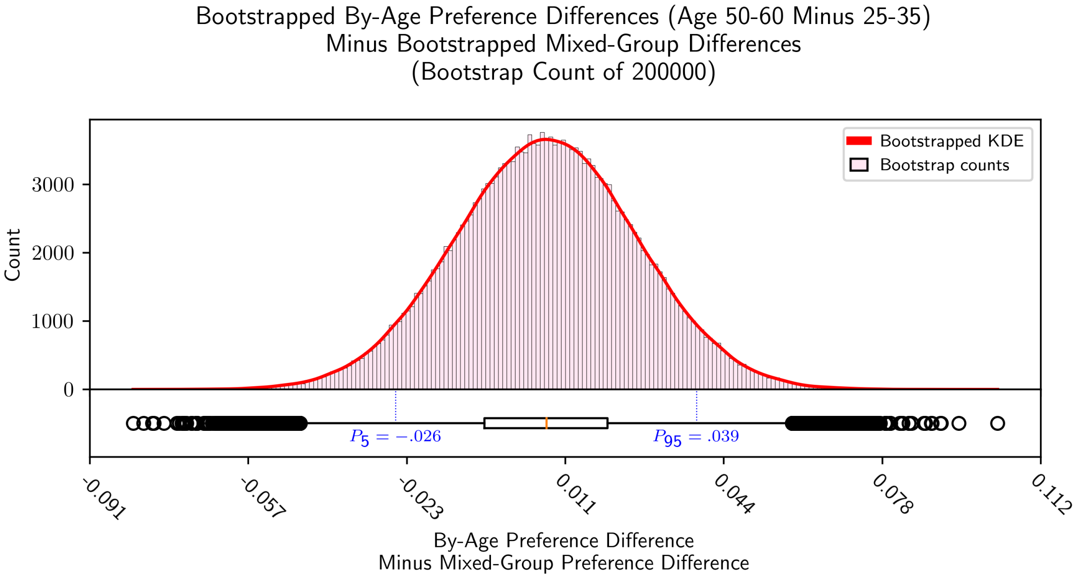
$\mathrm{CI}_{90} = [-0.086, -0.029]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

Dimension 12: valuable/special occasion-related



$p = 0.640095$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

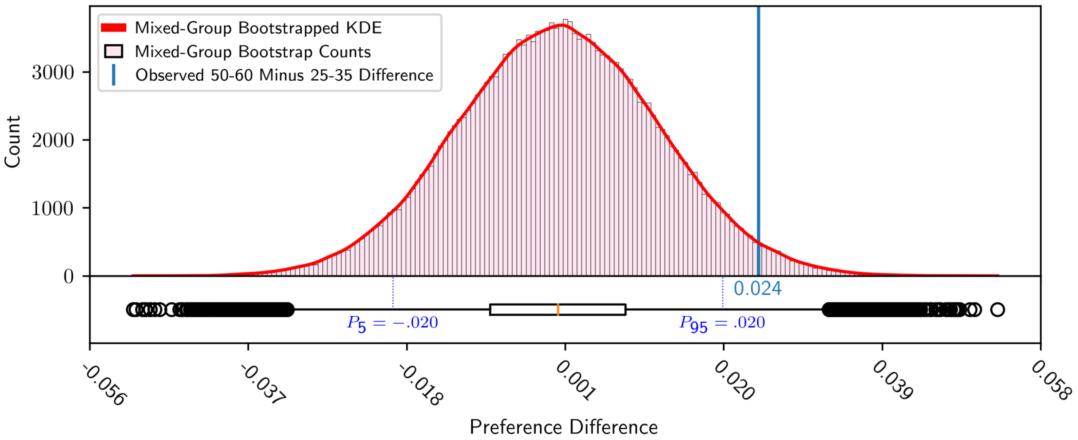
---



$\mathrm{CI}_{90} = [-0.026, 0.039]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

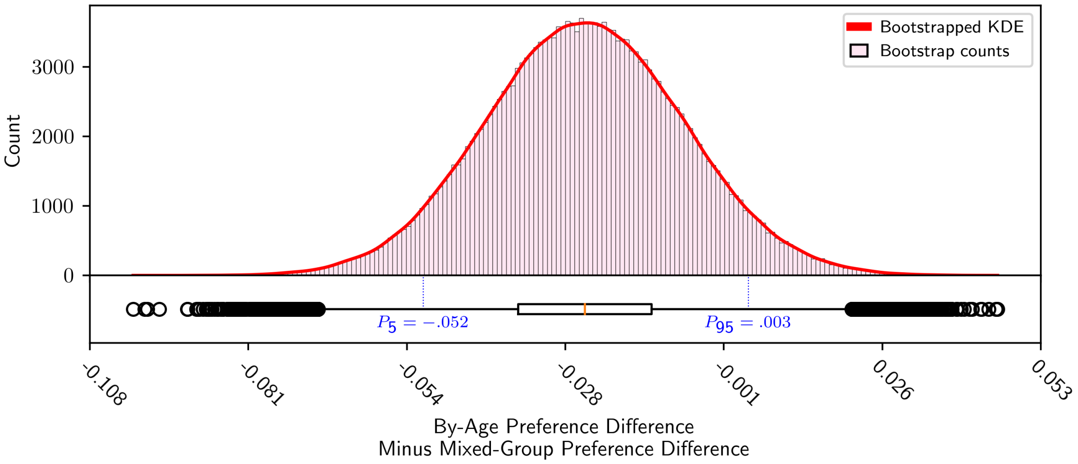
Dimension 13: electronic/technology

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.045095$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

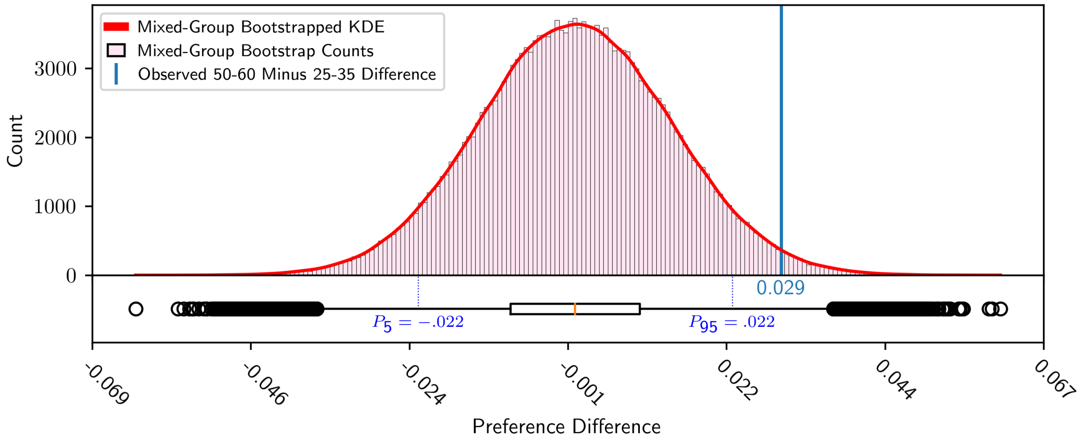
Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



$\mathrm{CI}_{90} = [-0.052, 0.003]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

Dimension 14: sport-related/recreational activity-related

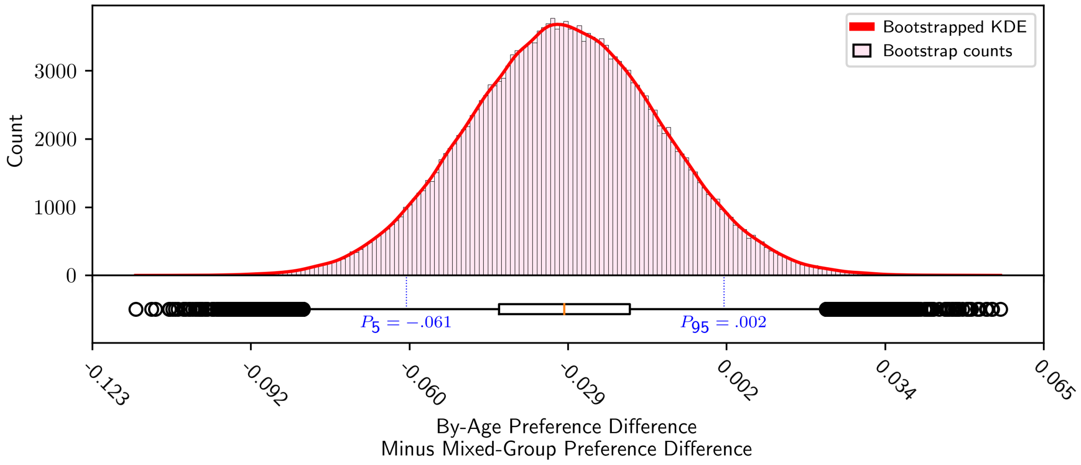
Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.03049$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

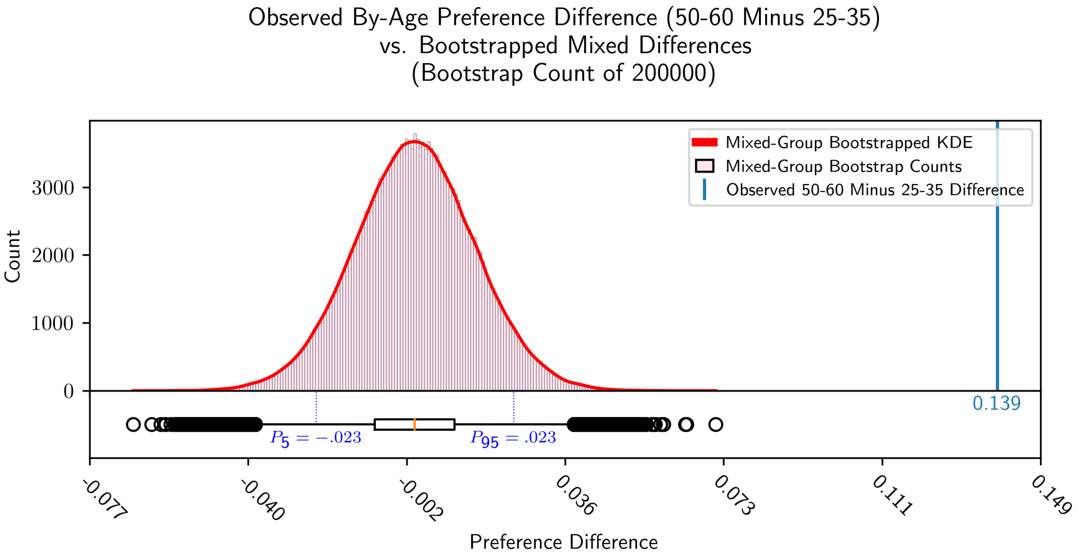
---

Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)

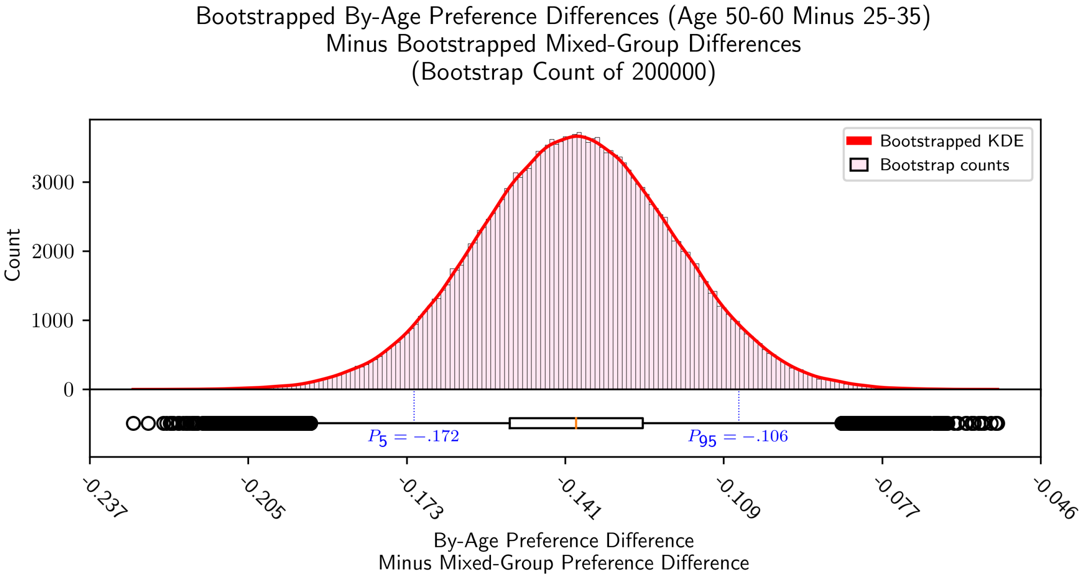


$\mathrm{CI}_{90} = [-0.061, 0.002]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

Dimension 15: disc-shaped/round

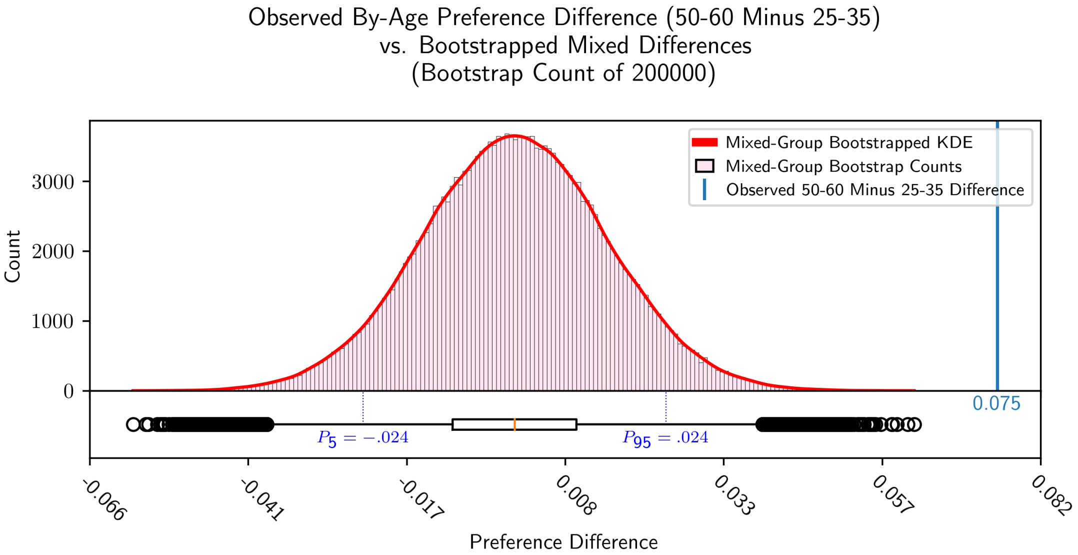


$p = 0.0$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

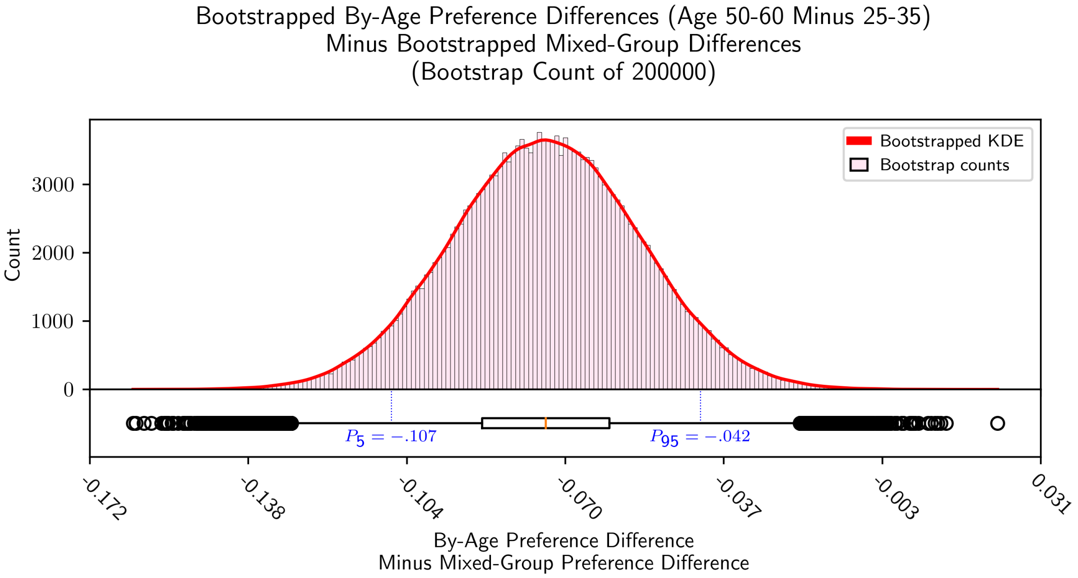


$\mathrm{CI}_{90} = [-0.172, -0.106]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

Dimension 16: tool-related

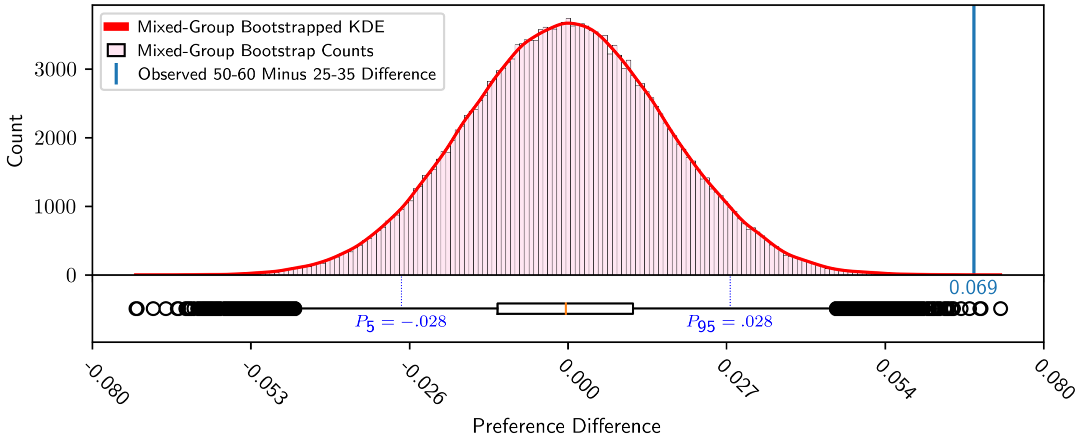


$p = 0.0$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution



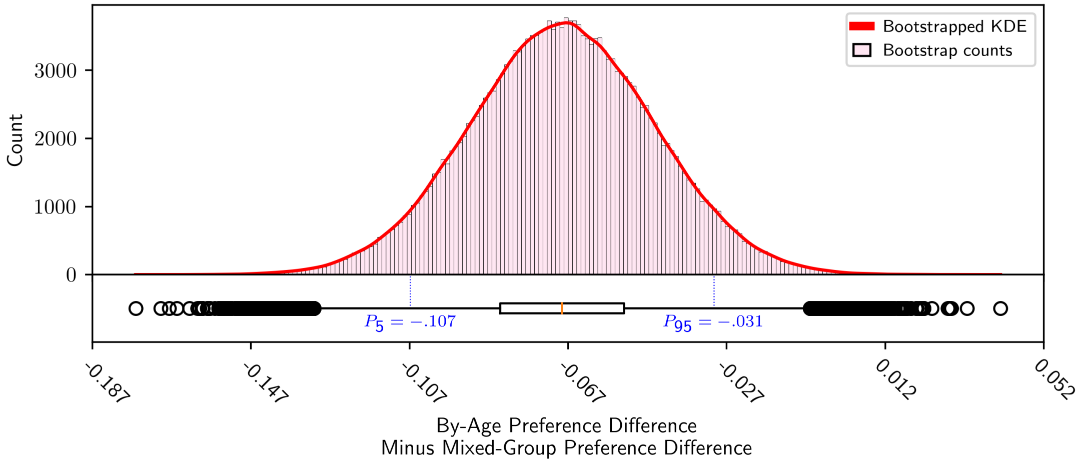
$\mathrm{CI}_{90} = [-0.107, -0.042]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

Dimension 17: many small things/course pattern



Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)

$p = 3e - 05$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

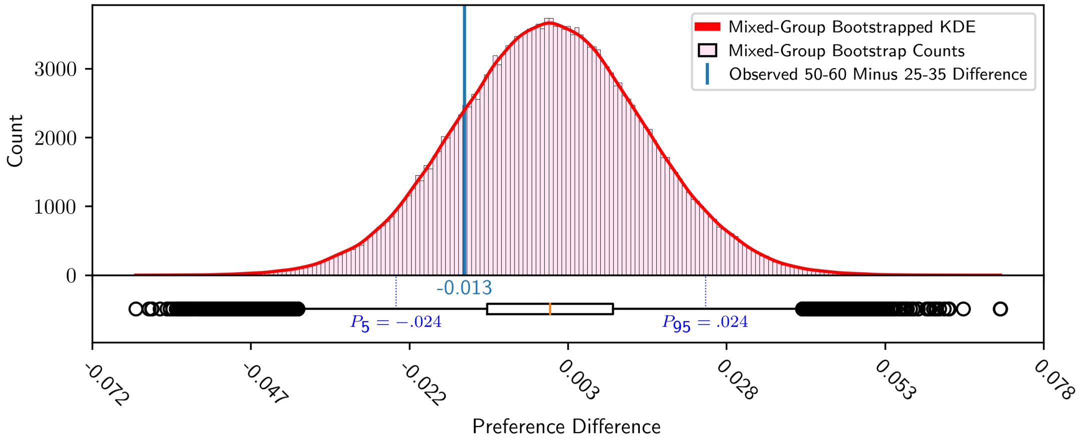


Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)

$\mathrm{CI}_{90} = [-0.107, -0.031]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

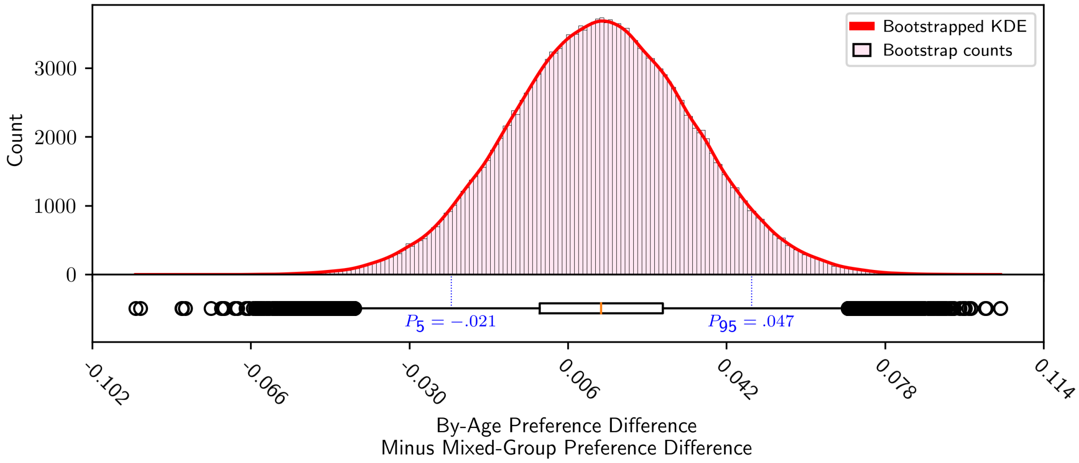
Dimension 18: paper-related/thin/flat/text-related

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.360775$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

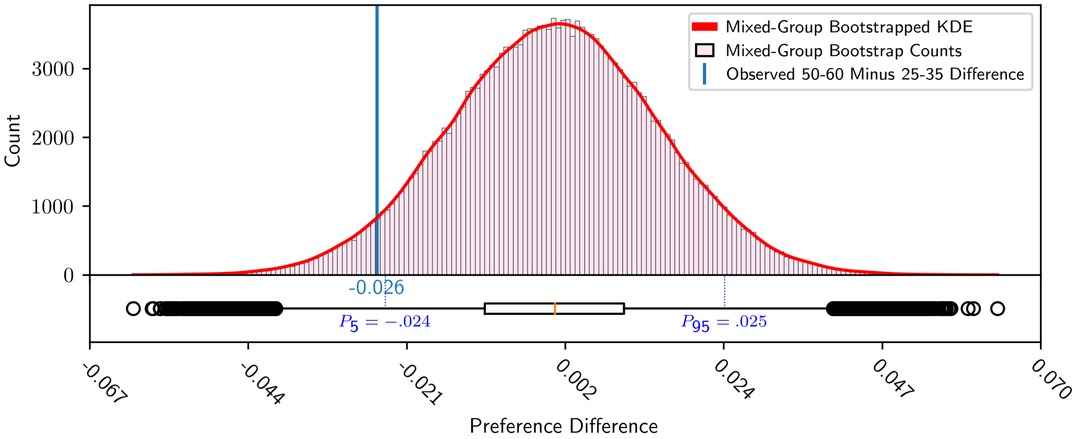
Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



$\mathrm{CI}_{90} = [-0.021, 0.047]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

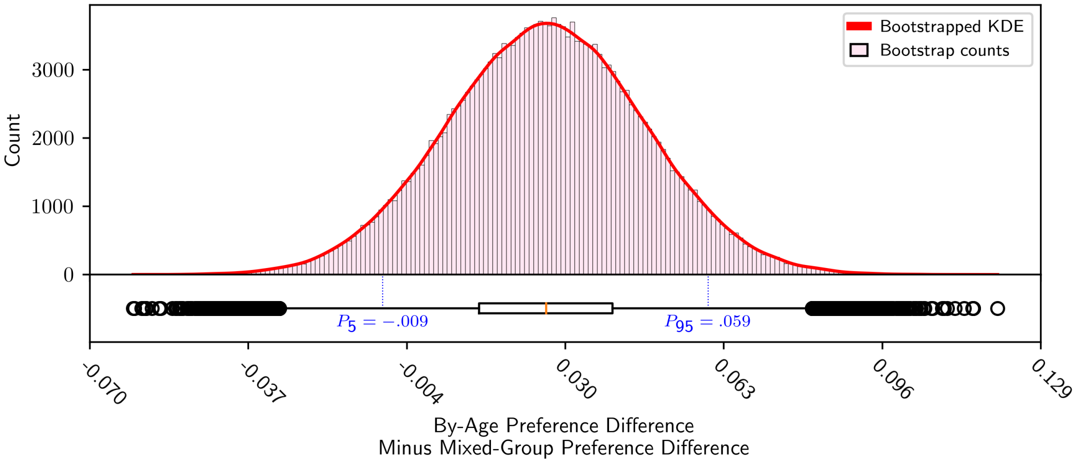
Dimension 19: fluid-related/drink-related

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



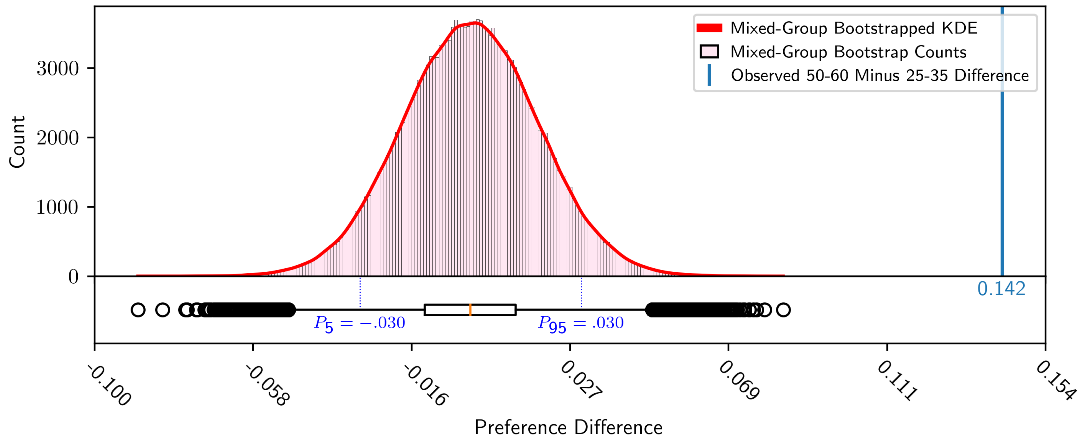
$p = 0.08548$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



$\mathrm{CI}_{90} = [-0.009, 0.059]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

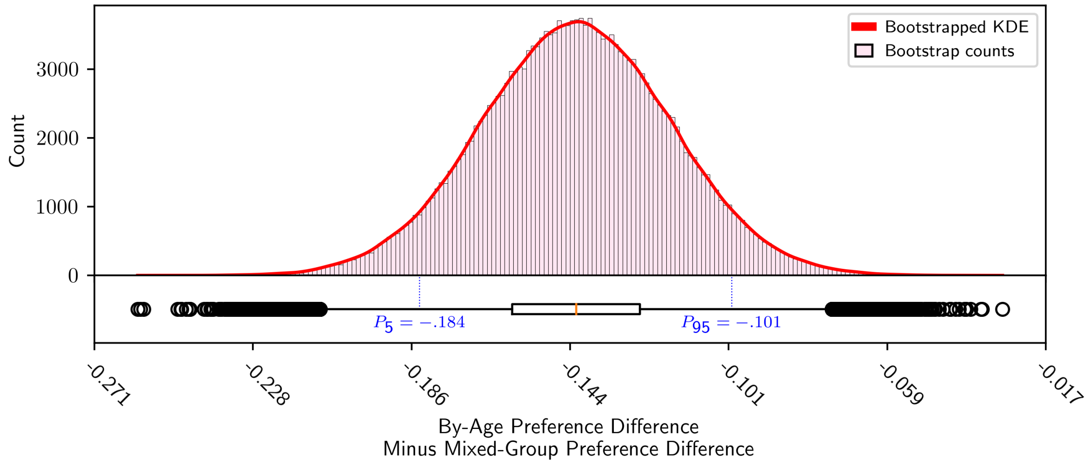
Dimension 20: long/thin



Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)

$p = 0.0$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

---

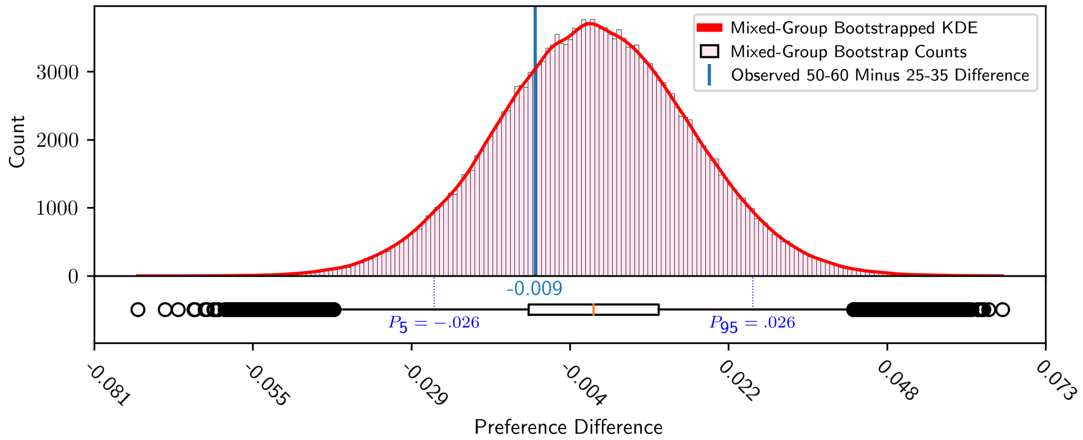


Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)

$\mathrm{CI}_{90} = [-0.184, -0.101]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

Dimension 21: water-related/blue

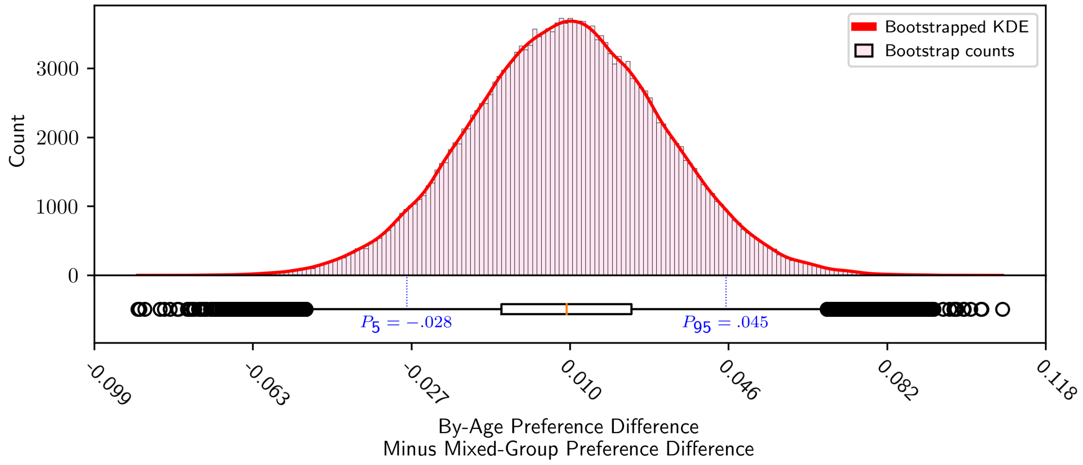
Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.54426$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

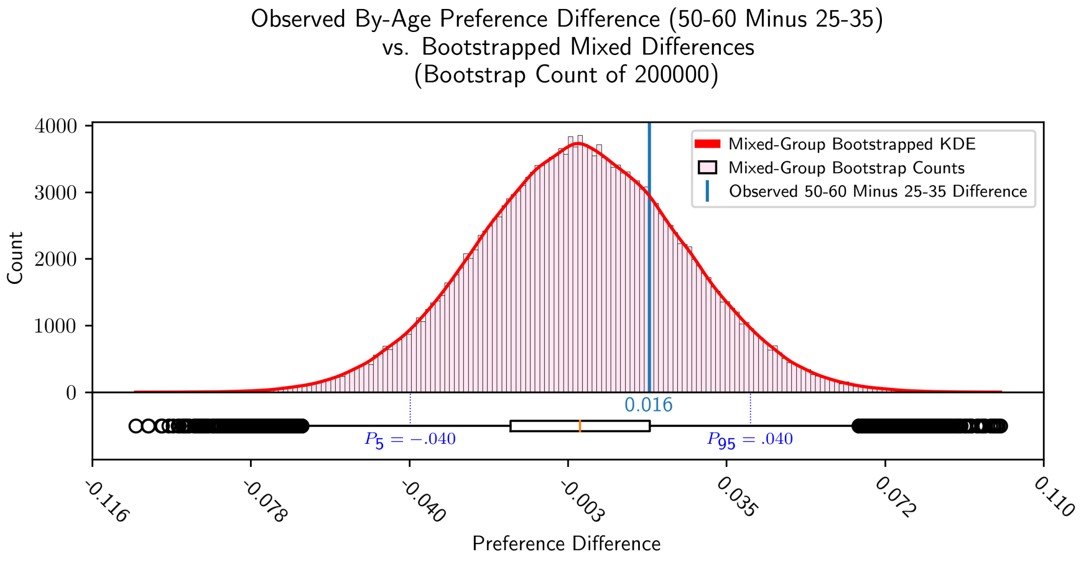
---

Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)

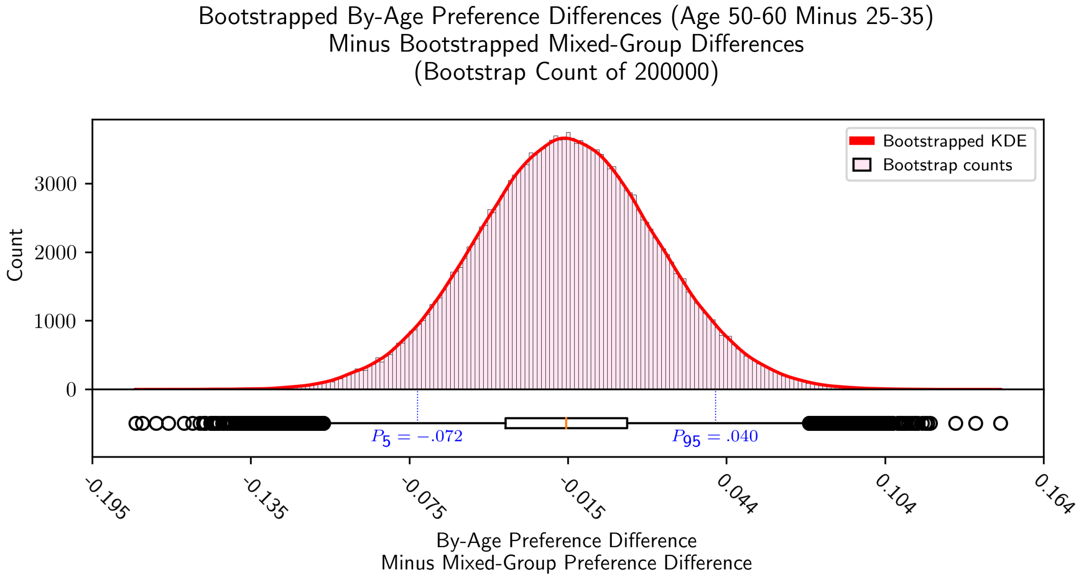


$\mathrm{CI}_{90} = [-0.028, 0.045]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

Dimension 22: powdery/fine-scale pattern

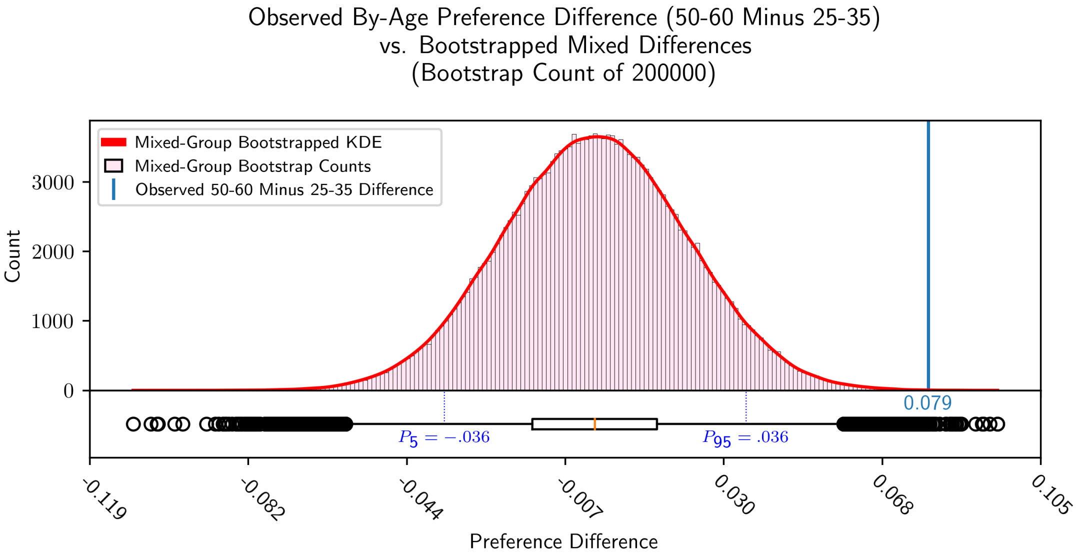


$p = 0.505935$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

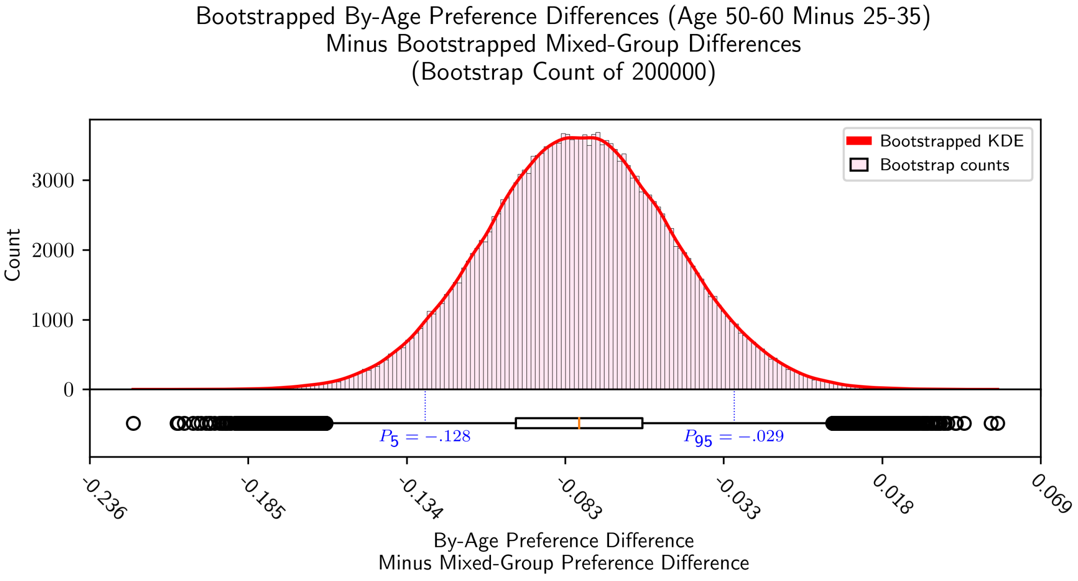


$\mathrm{CI}_{90} = [-0.072, 0.040]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

Dimension 23: red



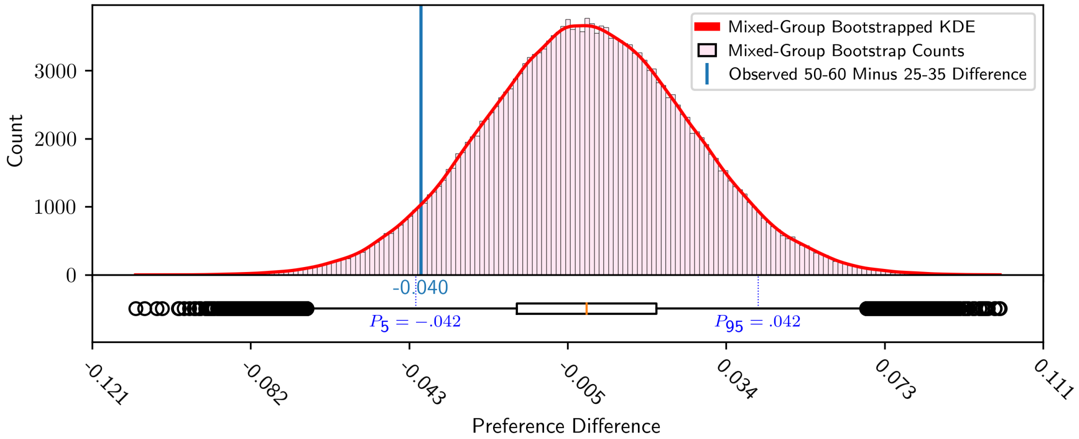
$p = 0.00032$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution



$\mathrm{CI}_{90} = [-0.128, -0.029]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

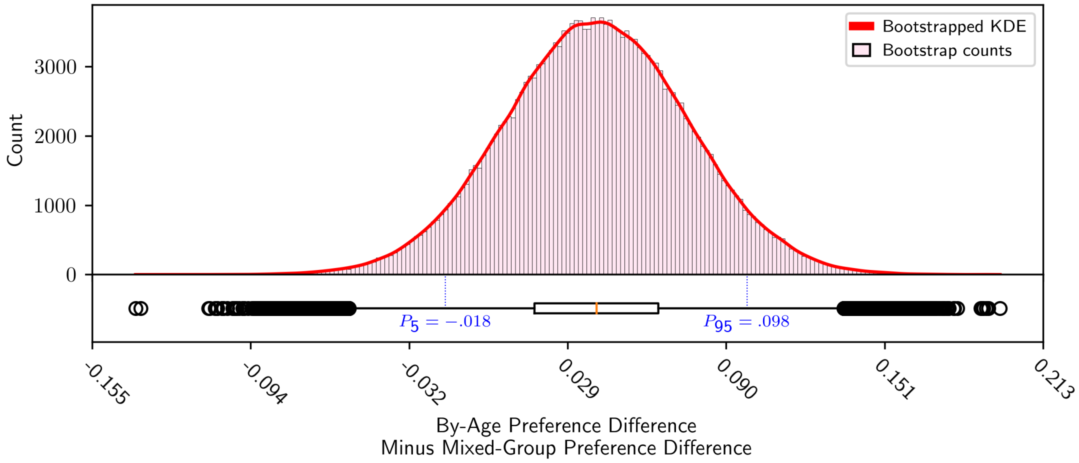
Dimension 24: feminine (stereotypically)/decorative

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.11074$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

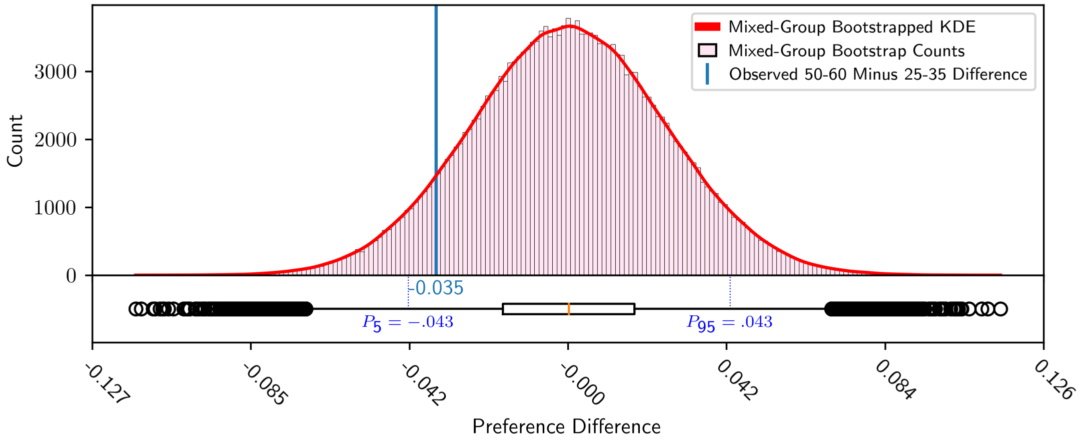
Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



$\mathrm{CI}_{90} = [-0.018, 0.098]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

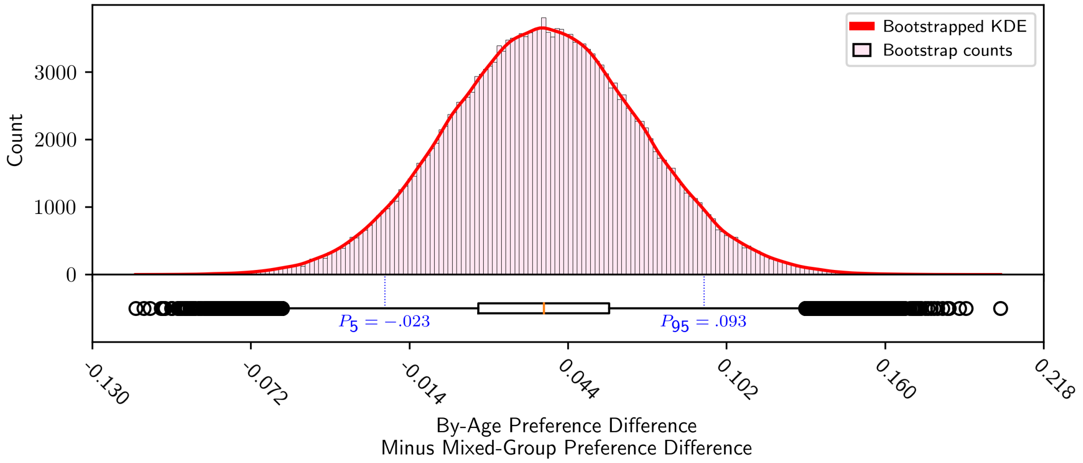
Dimension 25: bathroom-related/sanitary

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.174265$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

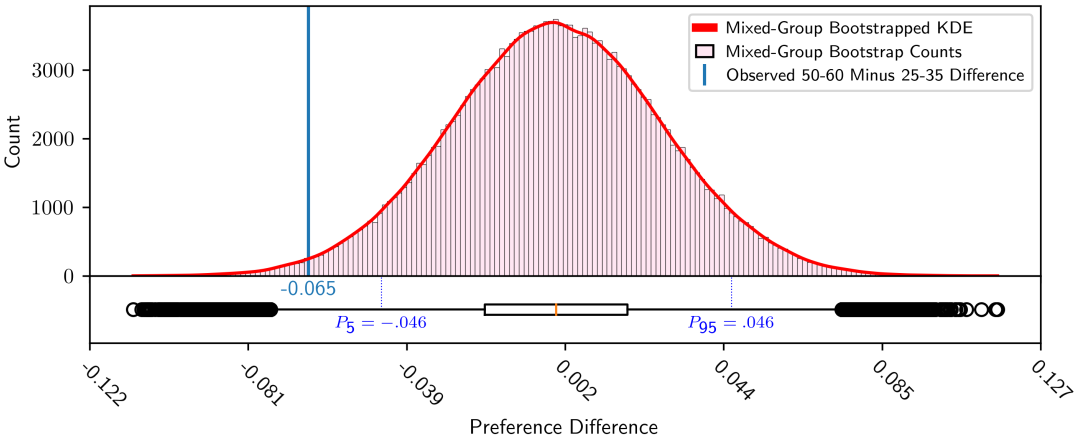
Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



$\mathrm{CI}_{90} = [-0.023, 0.093]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

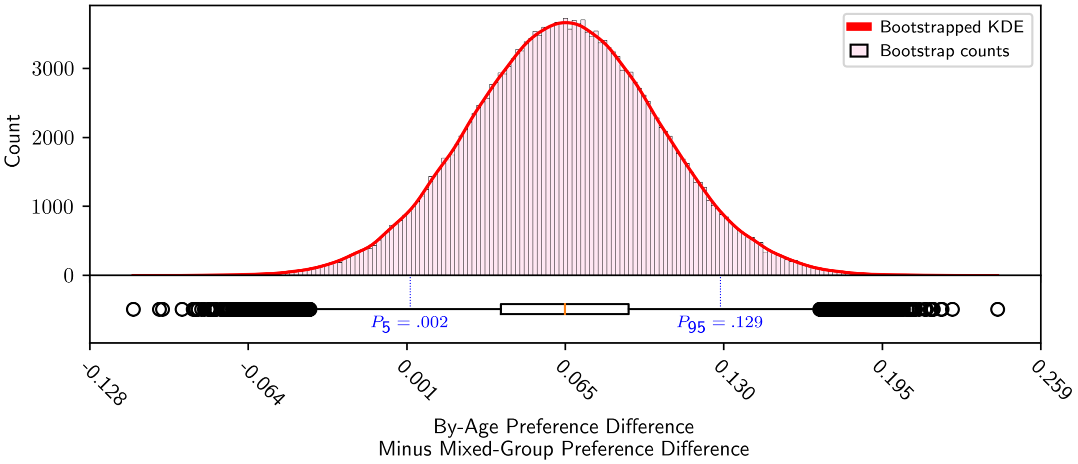
Dimension 26: black/noble

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



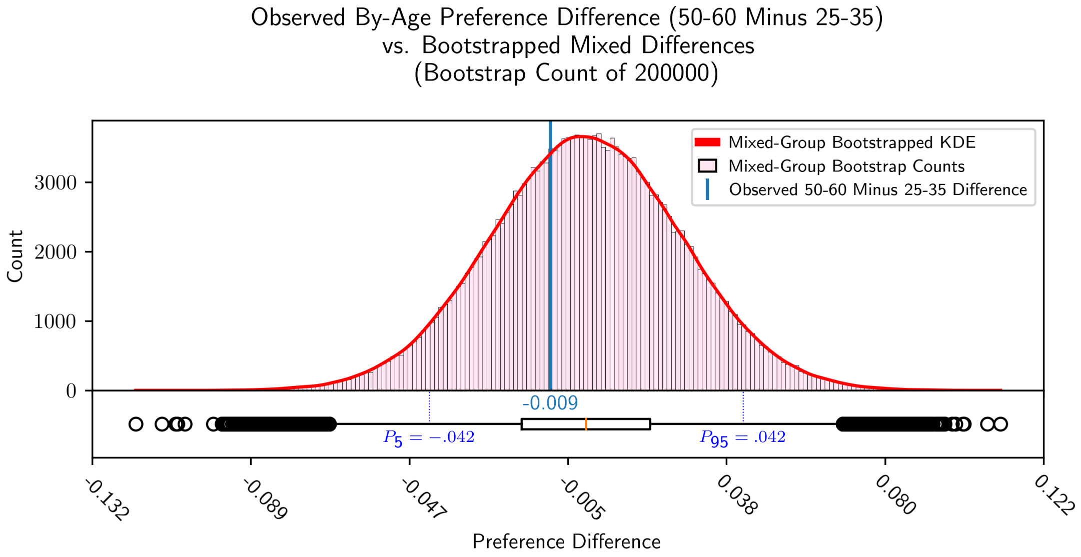
$p = 0.019695$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)

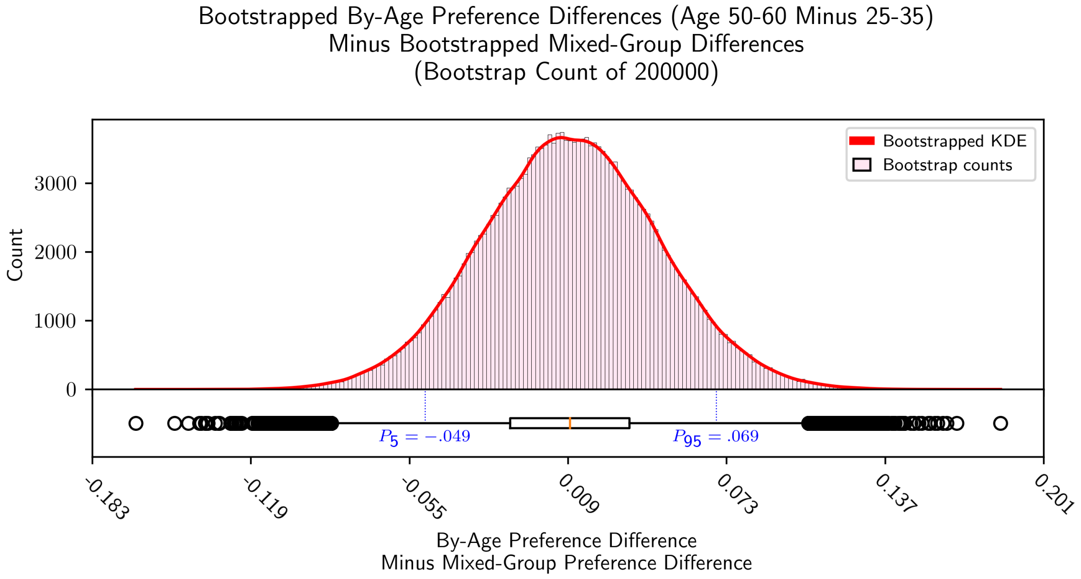


$\mathrm{CI}_{90} = [0.002, 0.129]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

Dimension 27: weapon/danger-related/violence

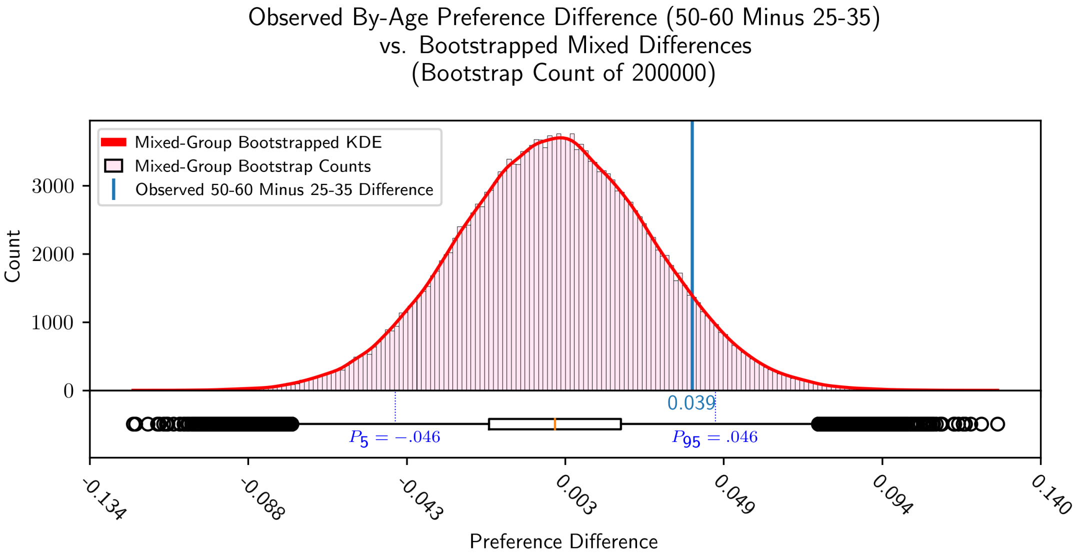


$p = 0.709965$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

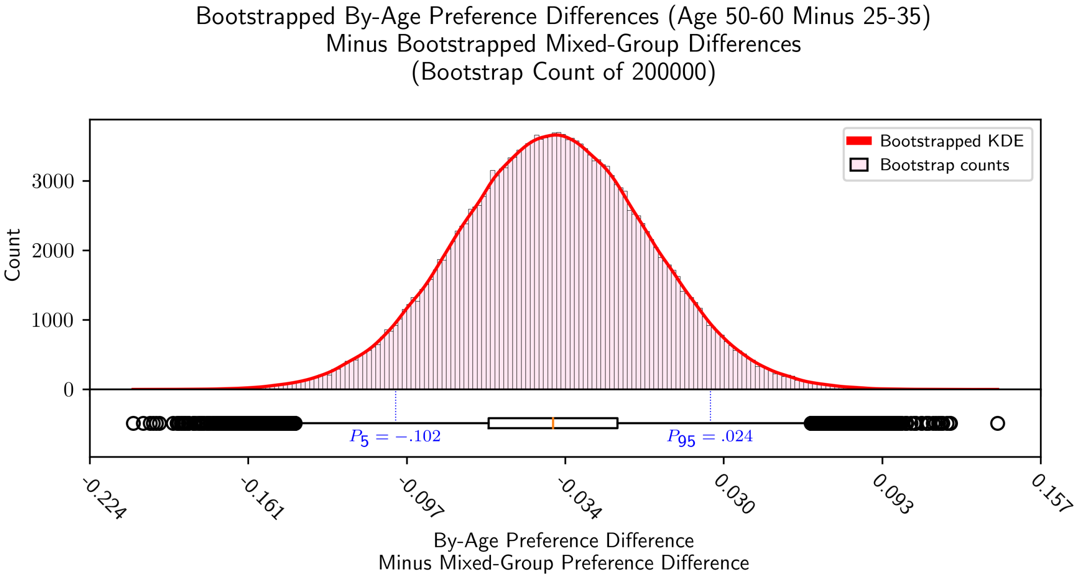


$\mathrm{CI}_{90} = [-0.049, 0.069]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

Dimension 28: musical instrument-related/noise-related



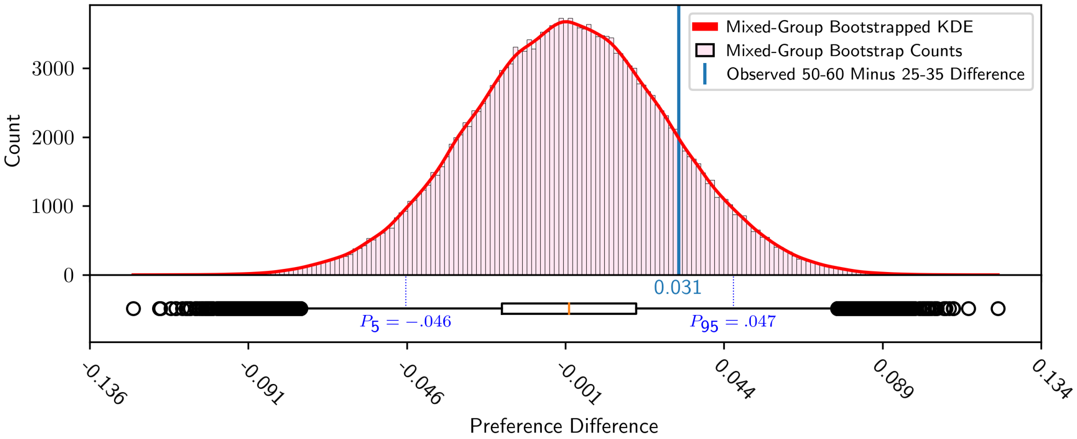
$p = 0.160265$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution



$\mathrm{CI}_{90} = [-0.102, 0.024]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

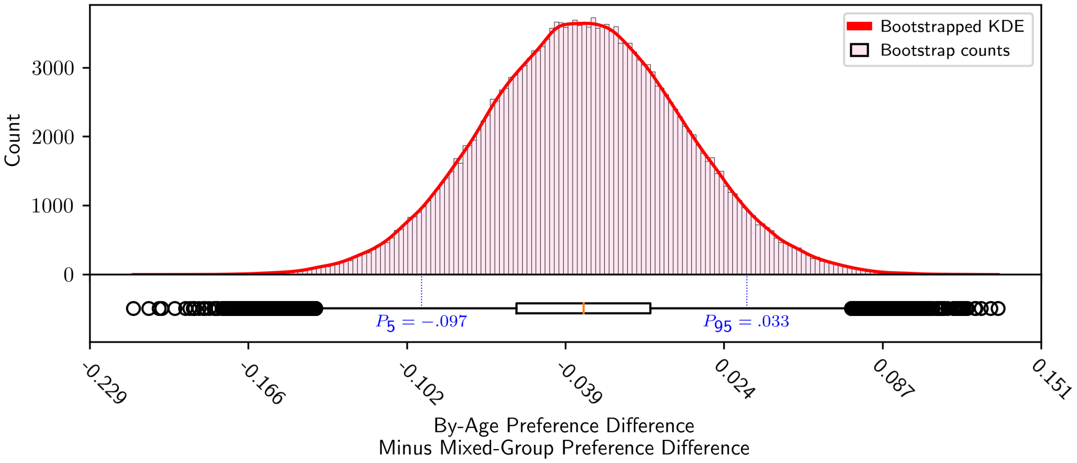
Dimension 29: sky-related/flying-related/floating-related

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.270265$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

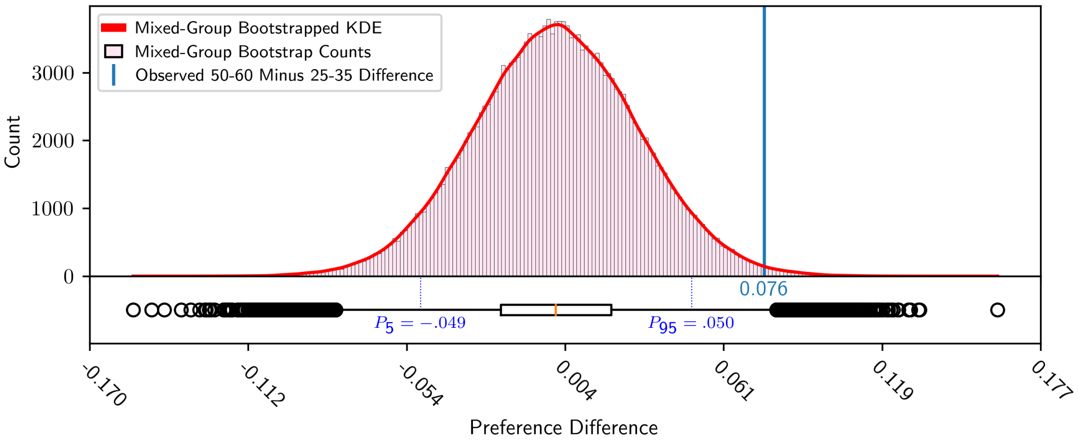
Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



$\mathrm{CI}_{90} = [-0.097, 0.033]$
for the distribution of subtraction of the difference between random groups from the difference between ages when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

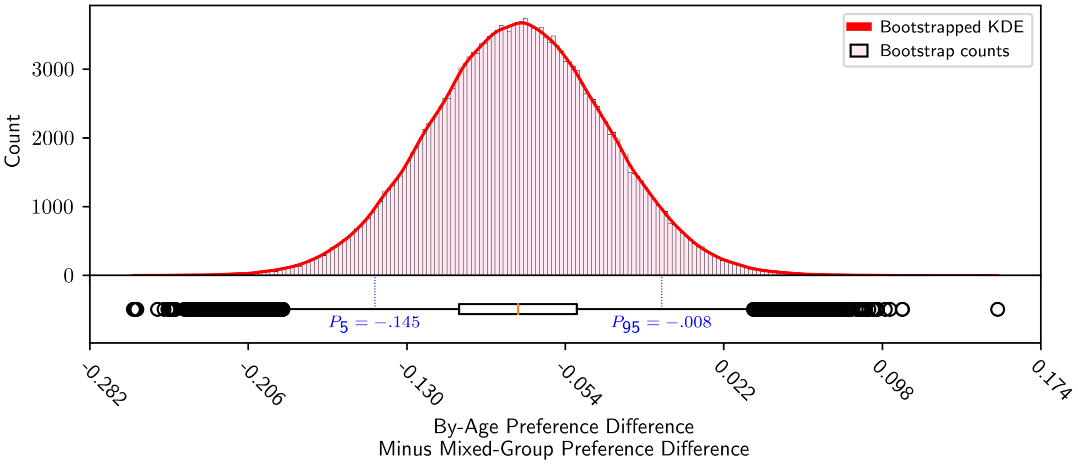
Dimension 30: spherical/ellipsoid/rounded/voluminous

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.012075$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

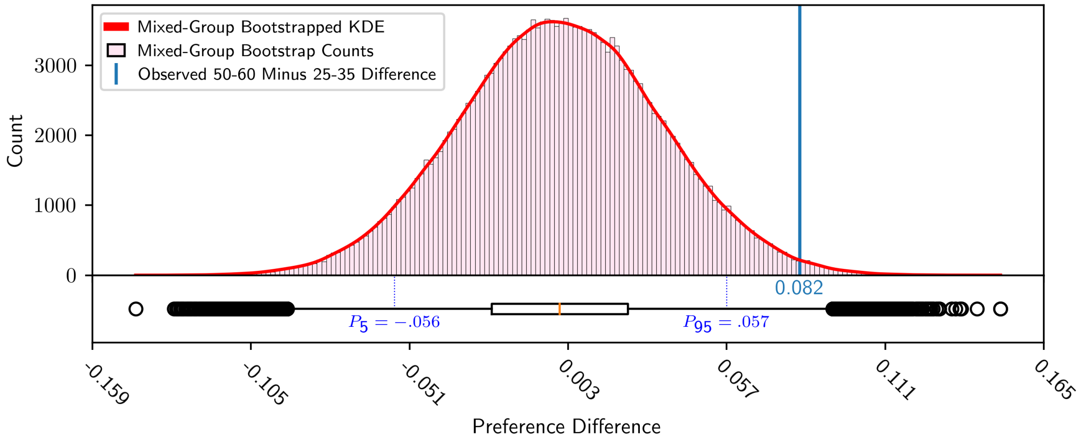
Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



$\mathrm{CI}_{90} = [-0.145, -0.008]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

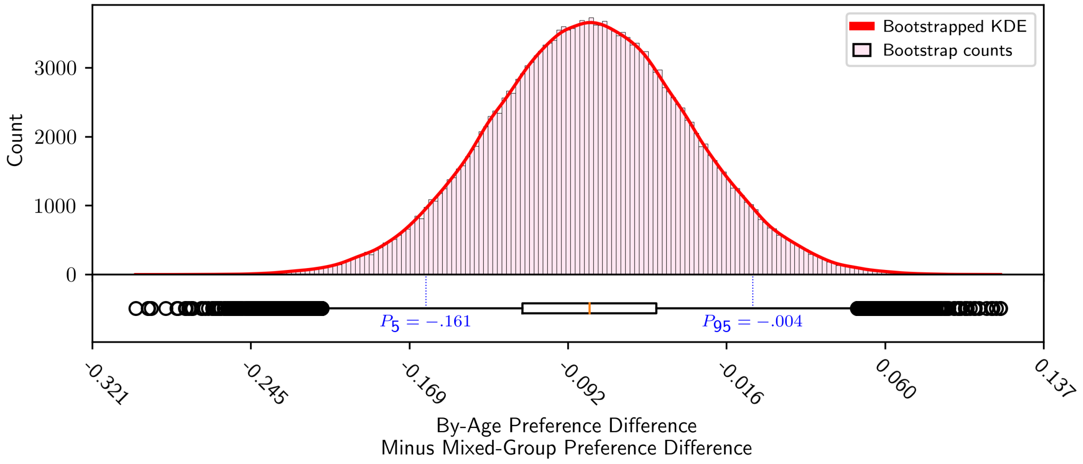
Dimension 31: repetitive

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.01651$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

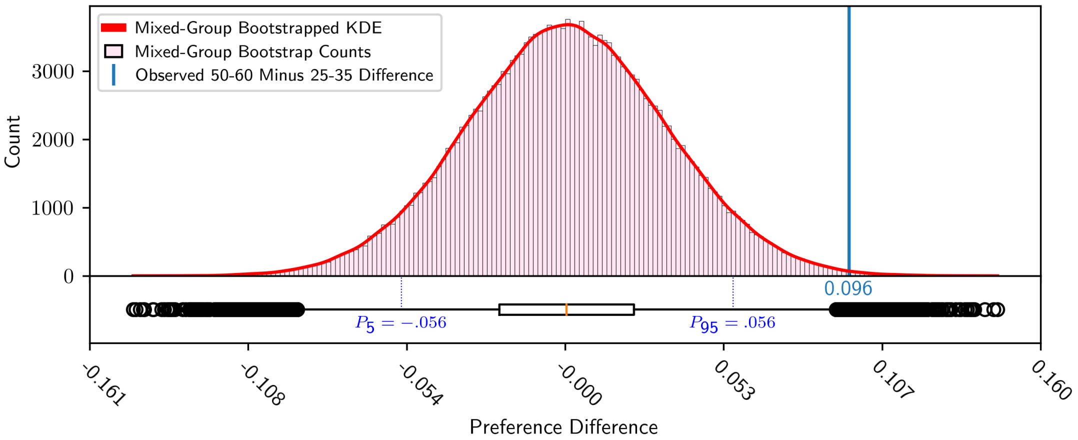
Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



$\mathrm{CI}_{90} = [-0.161, -0.004]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

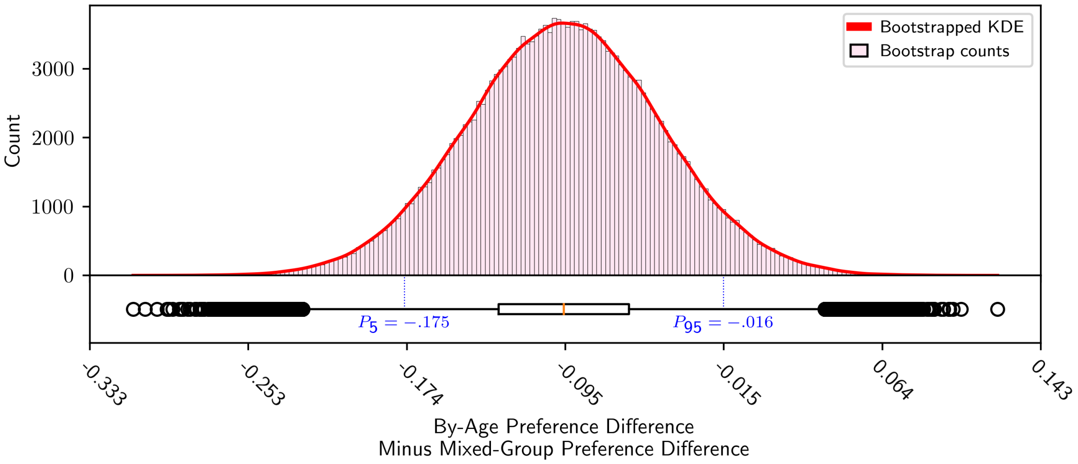
Dimension 32: flat/patterned

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



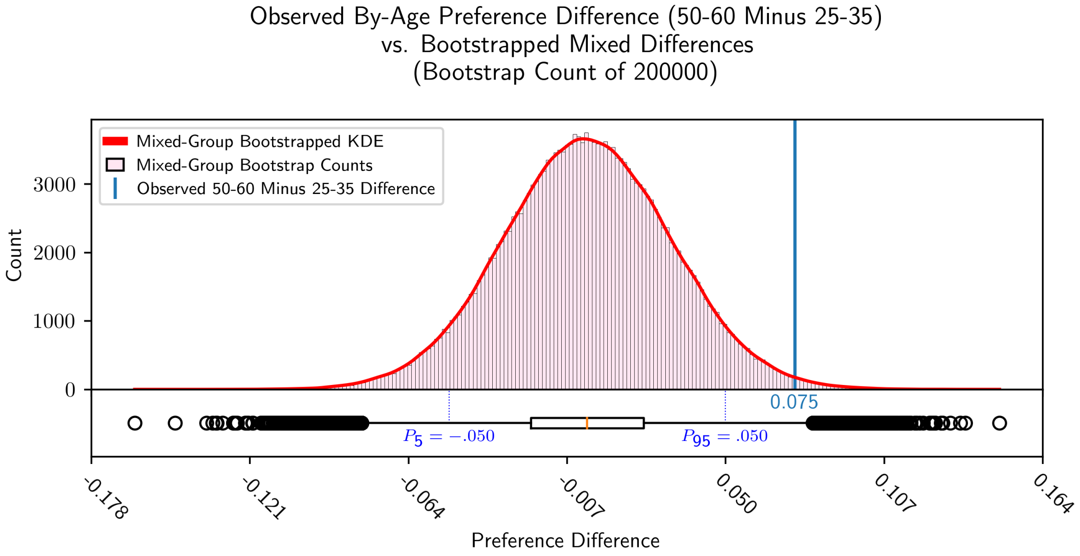
$p = 0.00495$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)

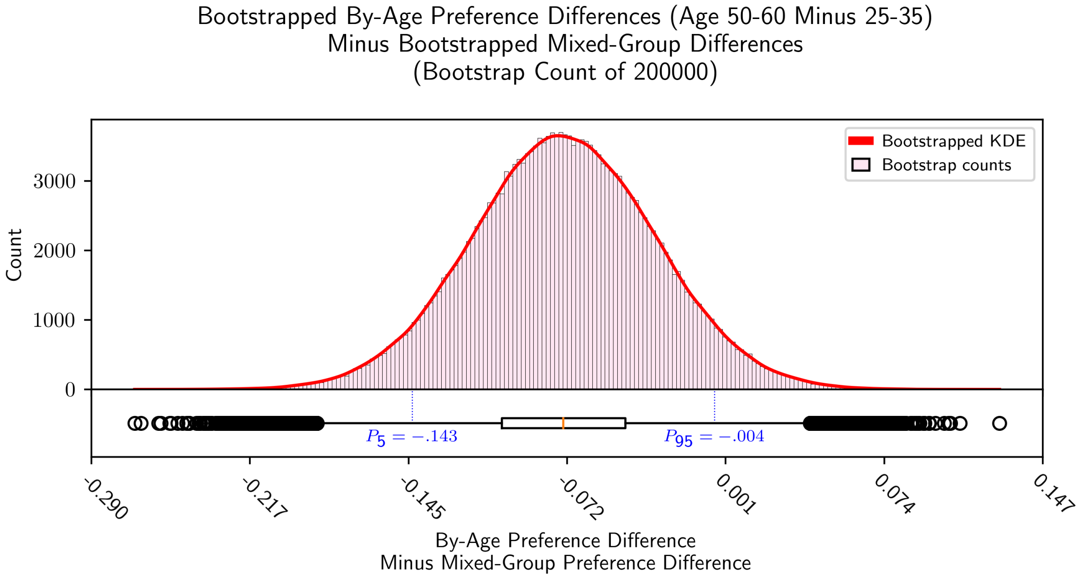


$\mathrm{CI}_{90} = [-0.175, -0.016]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

Dimension 33: white



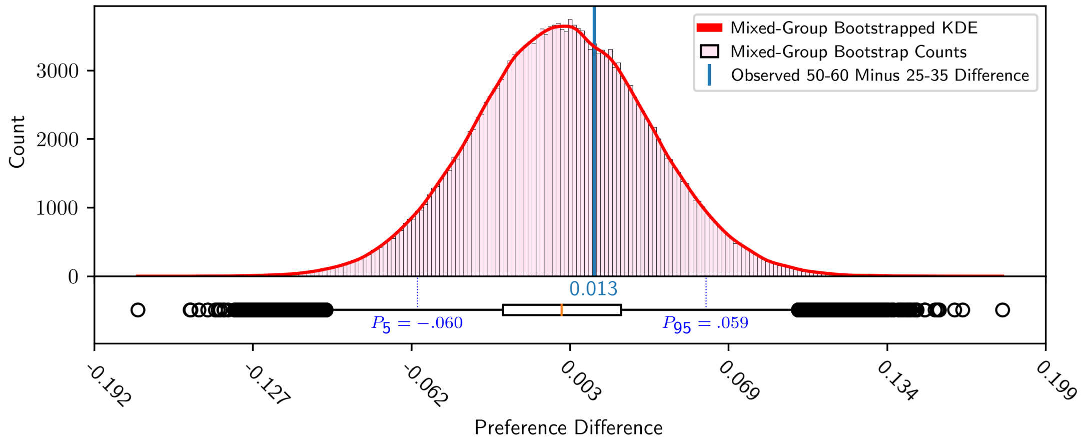
$p = 0.014625$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution



$\mathrm{CI}_{90} = [-0.143, -0.004]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

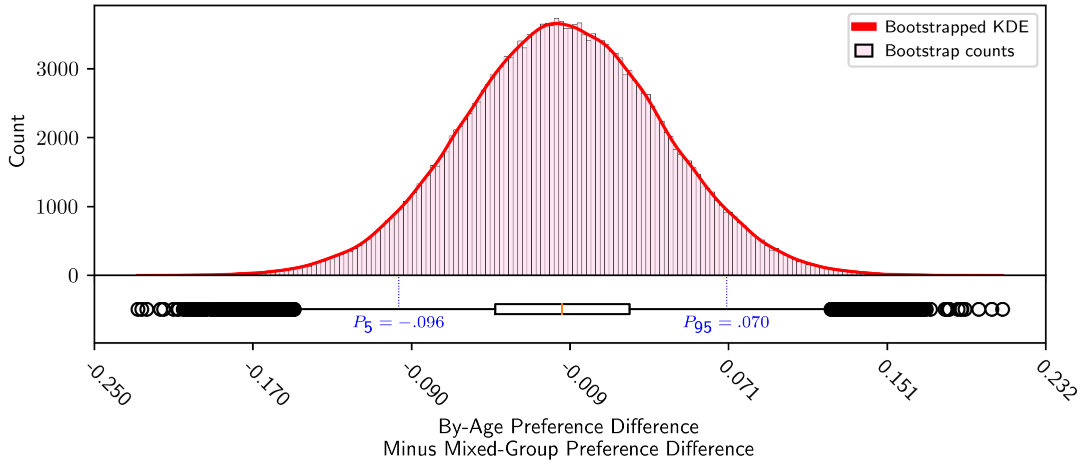
Dimension 34: thin/flat

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.71213$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

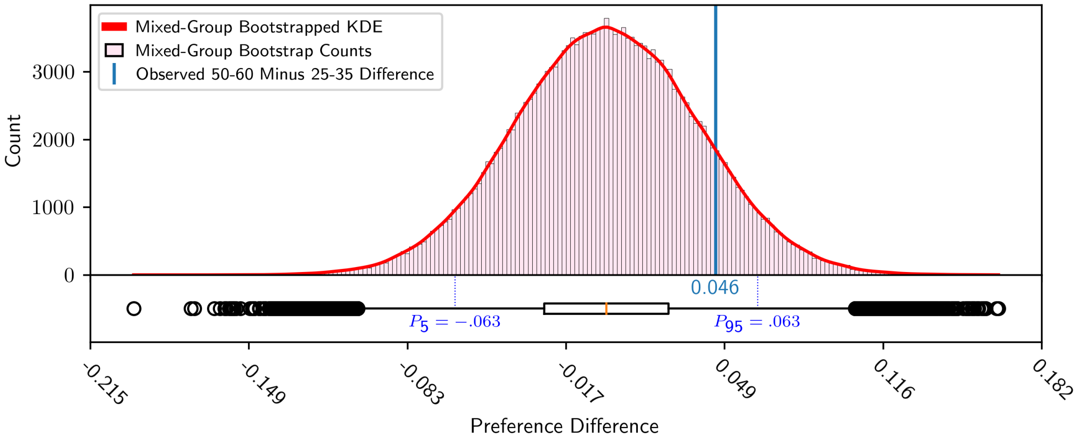
Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



$\mathrm{CI}_{90} = [-0.096, 0.070]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

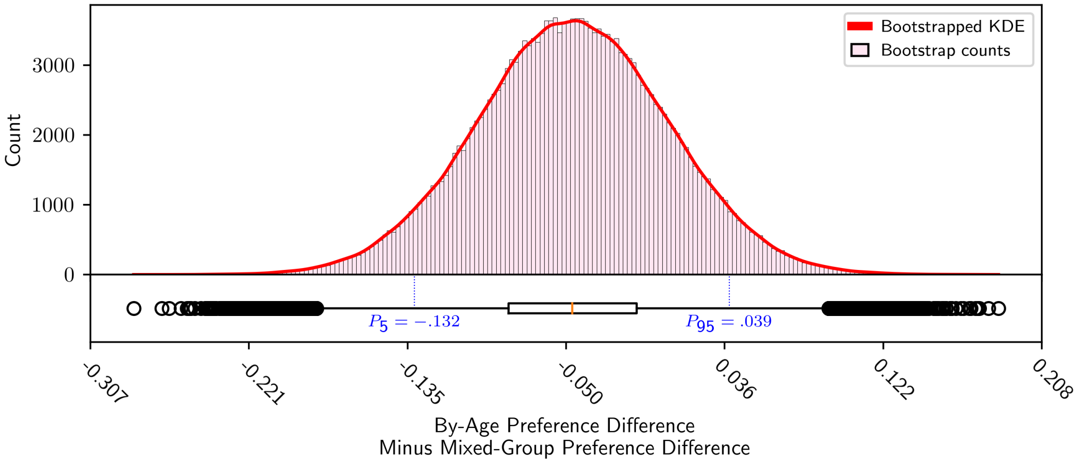
Dimension 35: disgusting/bugs

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



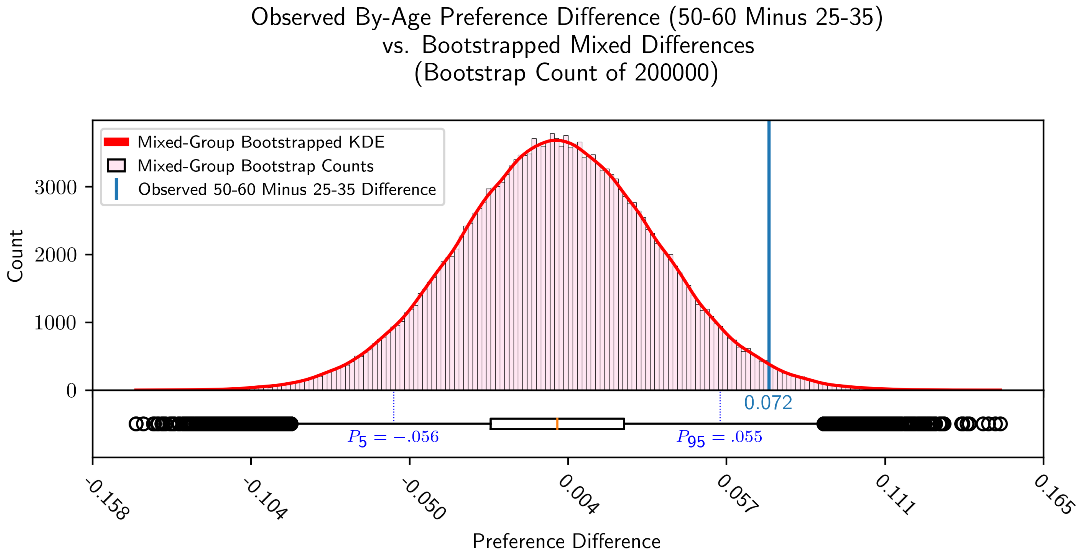
$p = 0.234525$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)

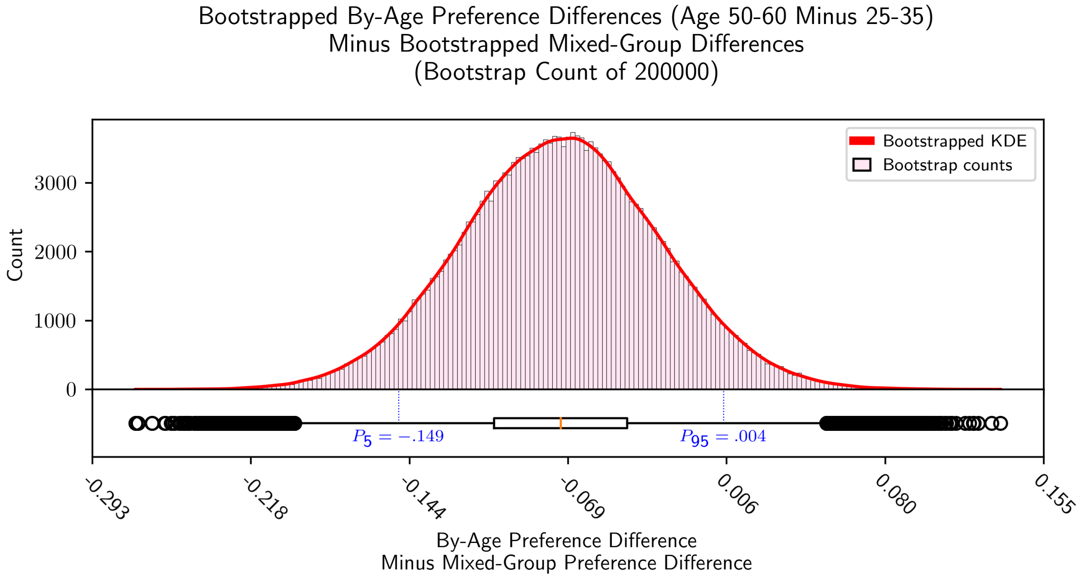


$\mathrm{CI}_{90} = [-0.132, 0.039]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

Dimension 36: string-related



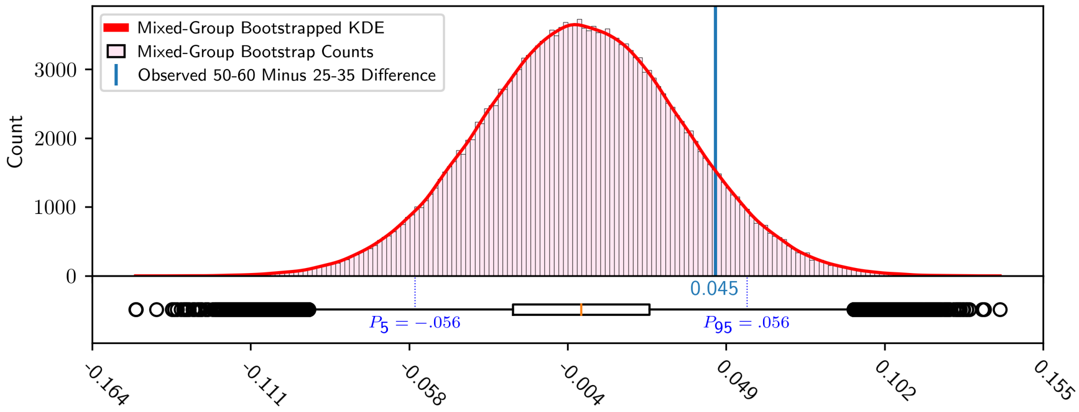
$p = 0.03368$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution



$\mathrm{CI}_{90} = [-0.149, 0.004]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

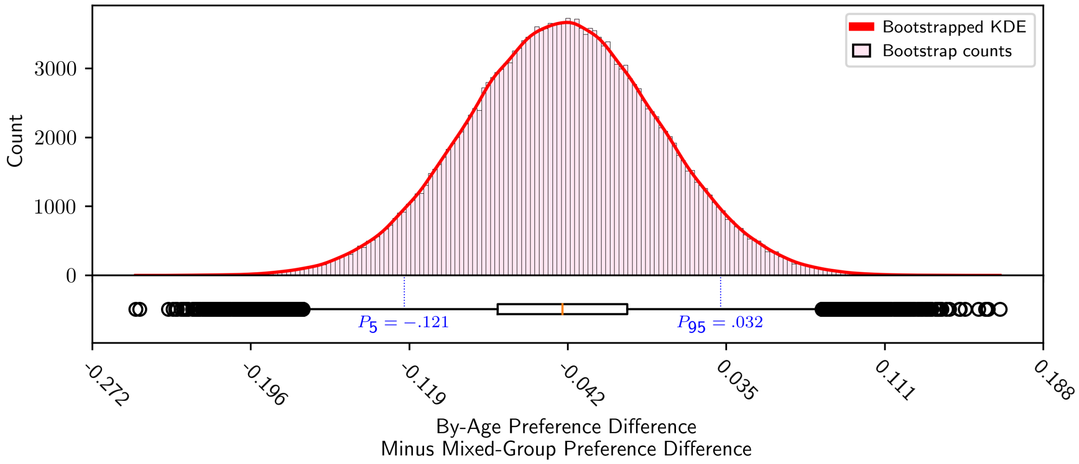
Dimension 37: arms/legs/skin-related

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.184315$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

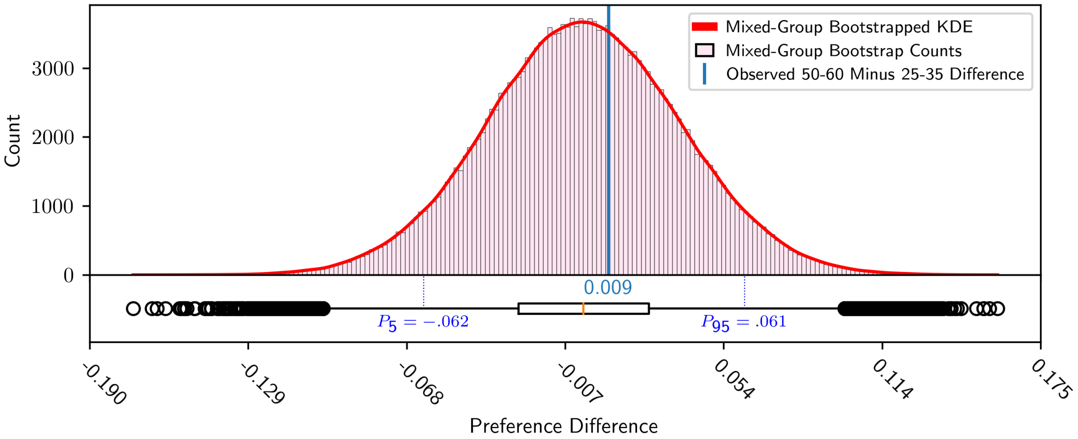
Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



$\mathrm{CI}_{90} = [-0.121, 0.032]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

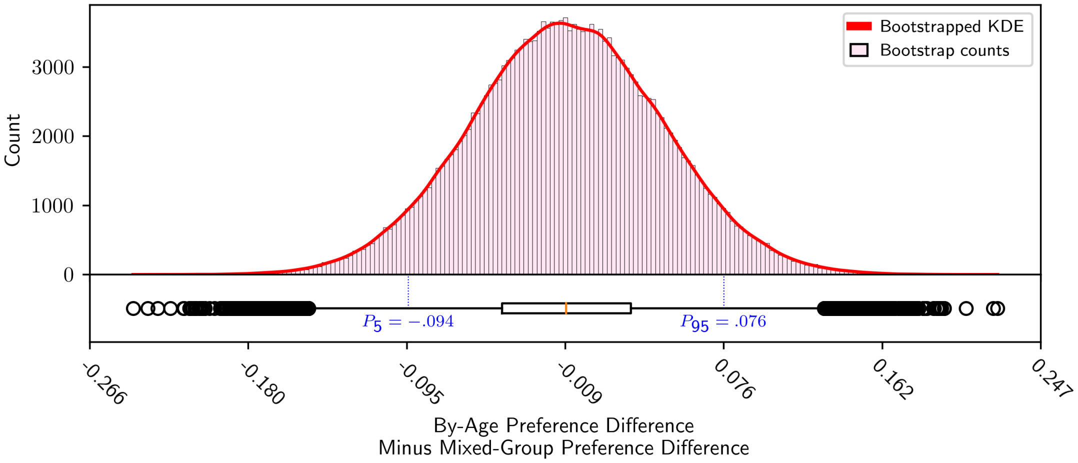
Dimension 38: shiny/transparent

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.79753$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

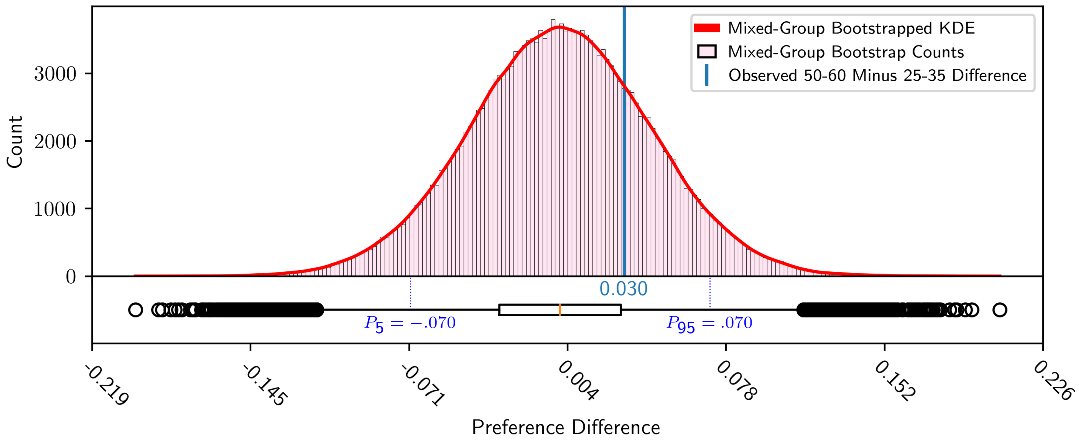
Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



$\mathrm{CI}_{90} = [-0.094, 0.076]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

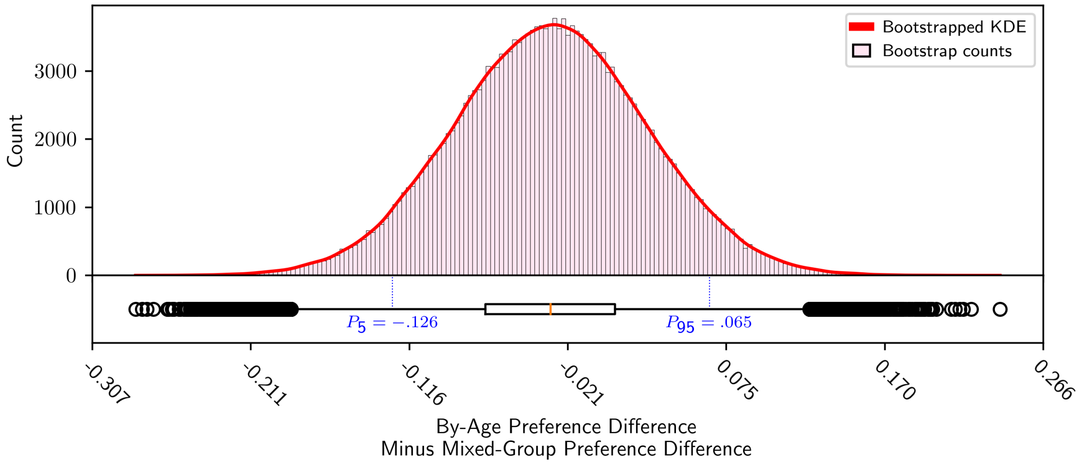
Dimension 39: construction-related/physical work-related

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.47505$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

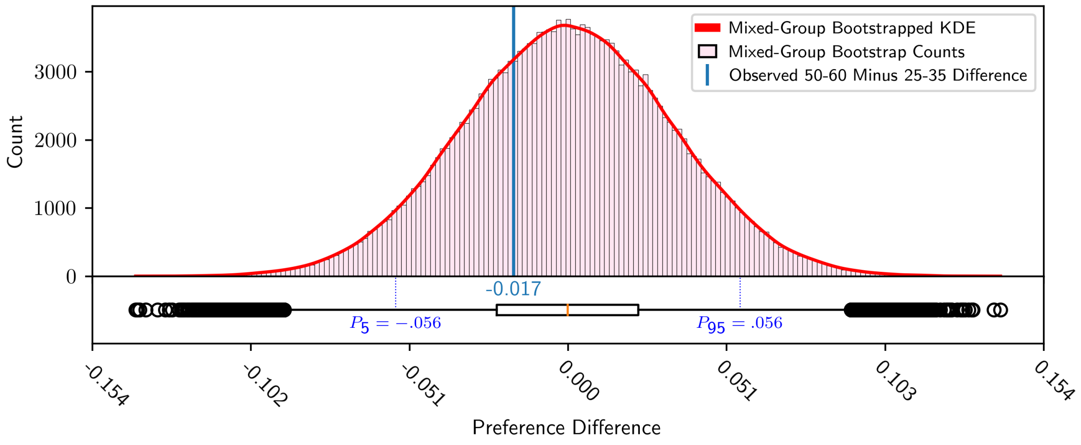
Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



$\mathrm{CI}_{90} = [-0.126, 0.065]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

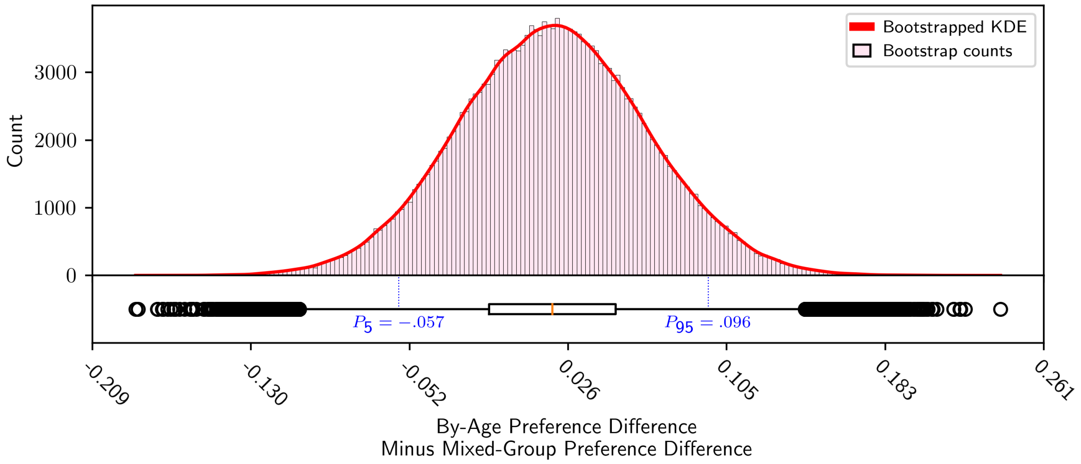
Dimension 40: fire-related/heat-related

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.60931$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

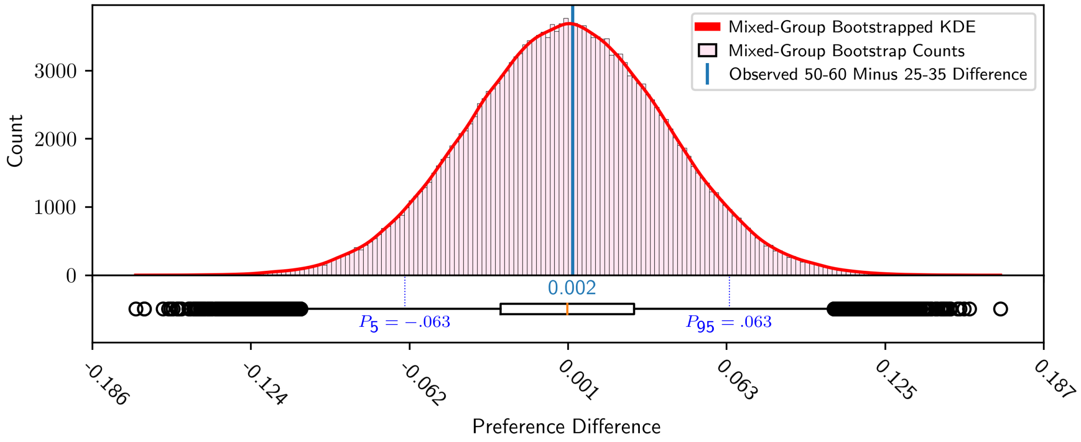
Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



$\mathrm{CI}_{90} = [-0.057, 0.096]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

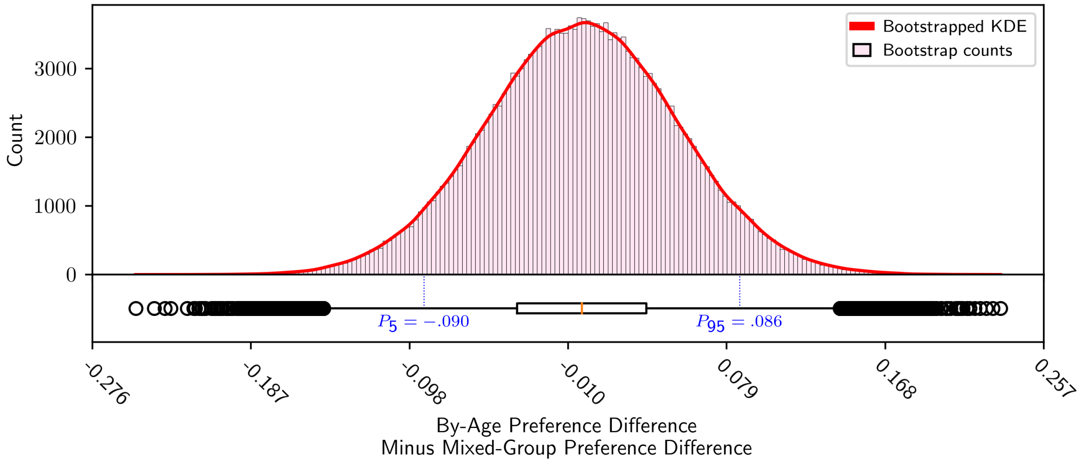
Dimension 41: head-related/face-related

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.95163$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

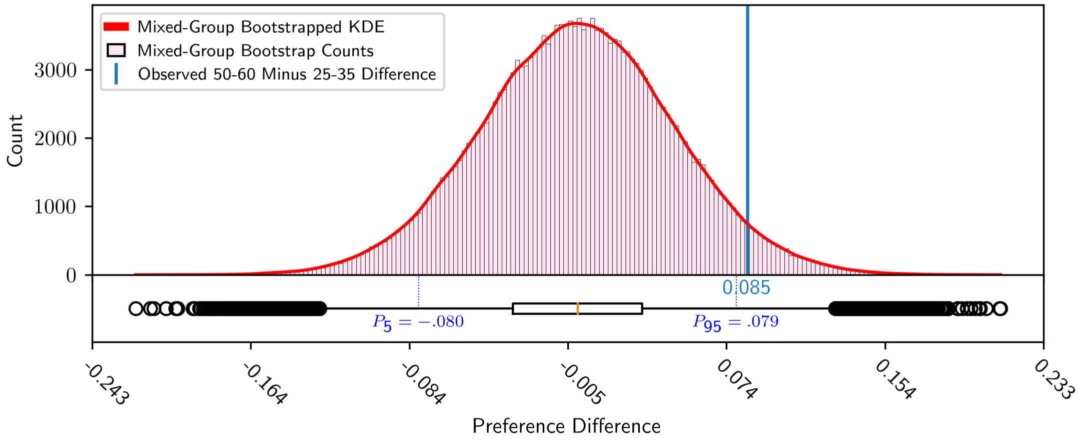
Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



$\mathrm{CI}_{90} = [-0.090, 0.086]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

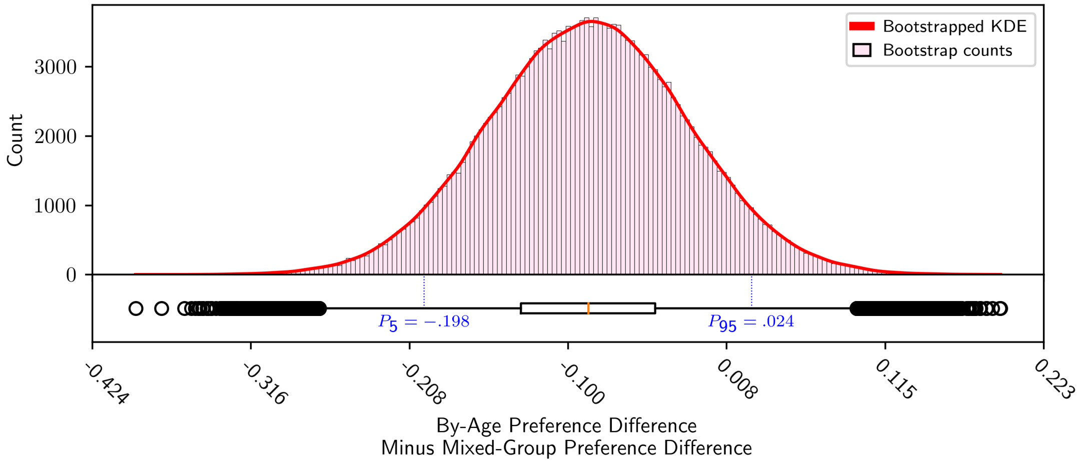
Dimension 42: beams-related

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



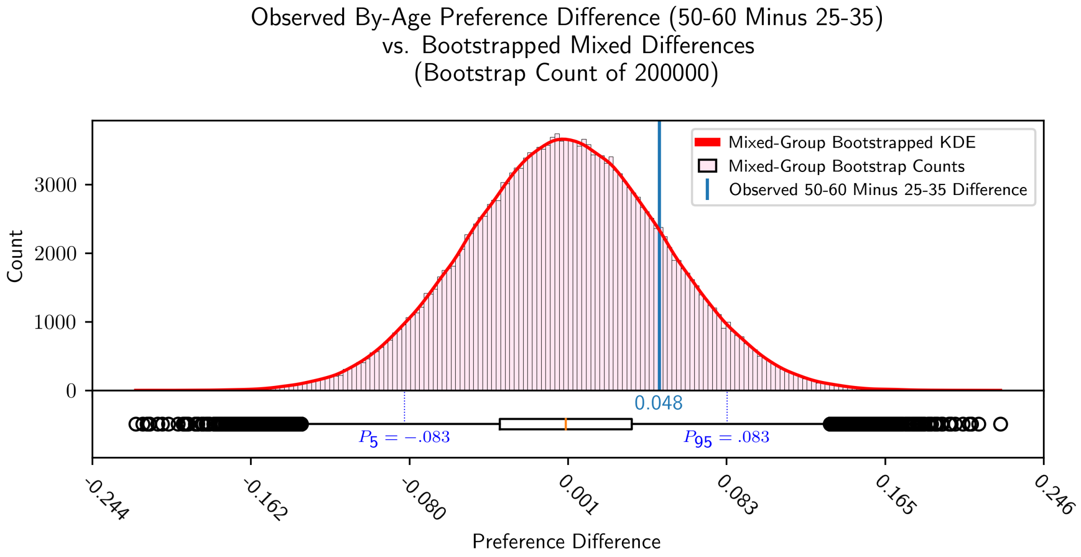
$p = 0.07992$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)

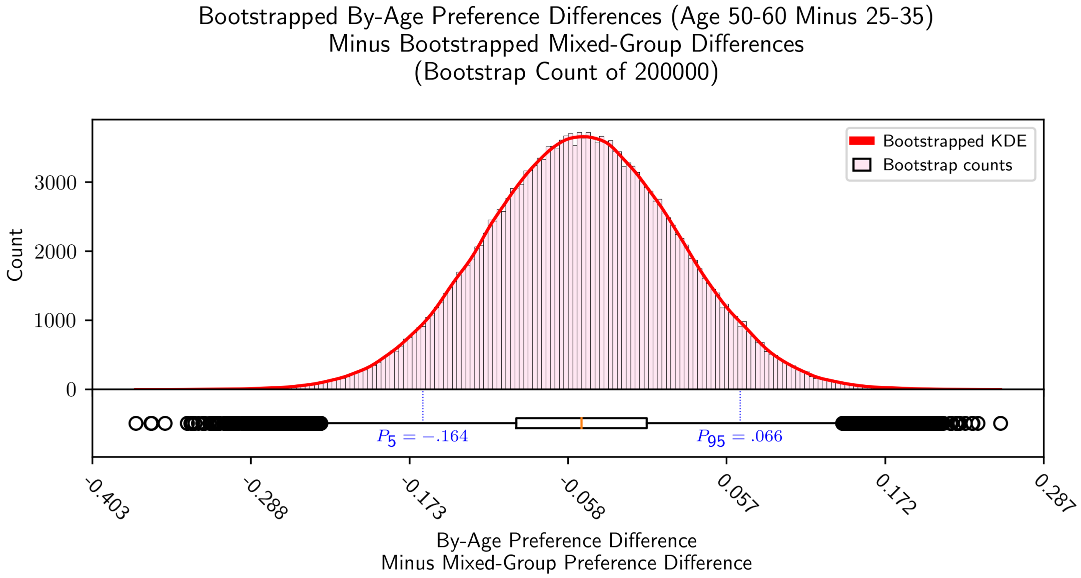


$\mathrm{CI}_{90} = [-0.198, 0.024]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

Dimension 43: seating-related/put things on top



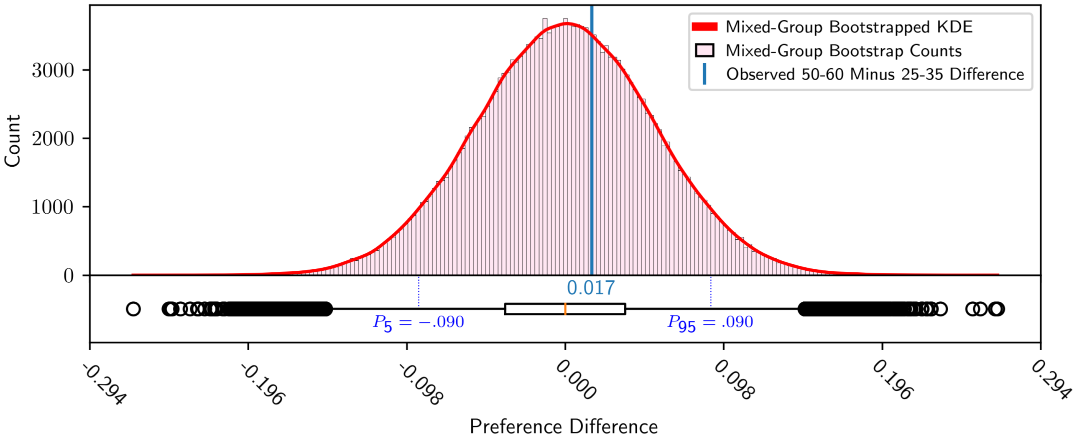
$p = 0.338365$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution



$\mathrm{CI}_{90} = [-0.164, 0.066]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

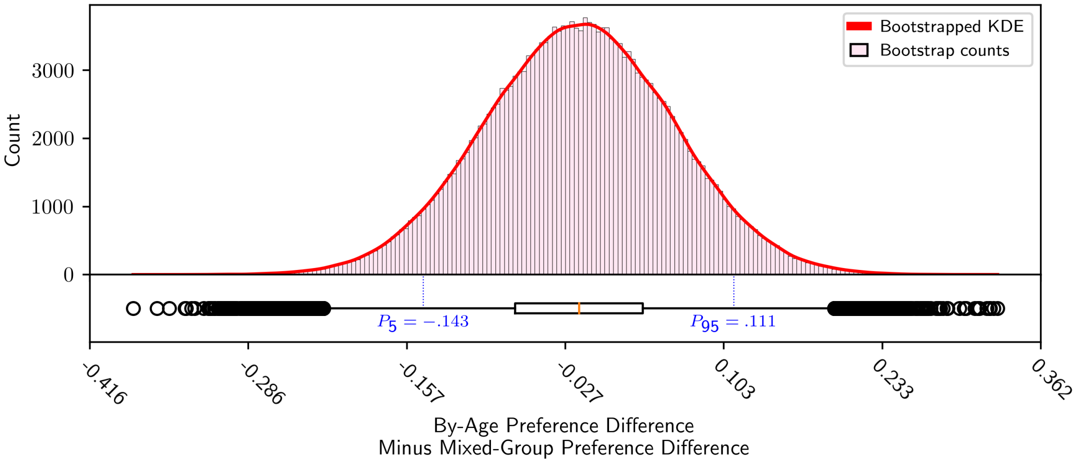
Dimension 44: container-related/hollow

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.76201$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

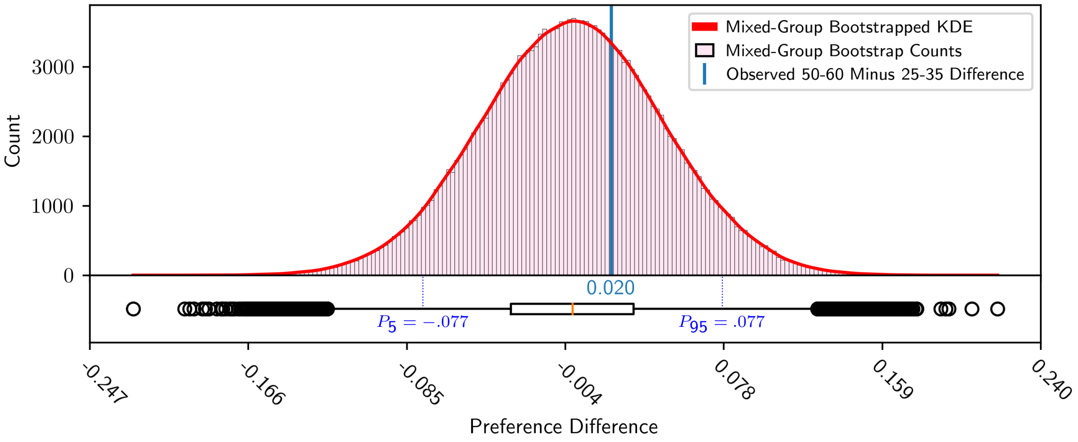
Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



$\mathrm{CI}_{90} = [-0.143, 0.111]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

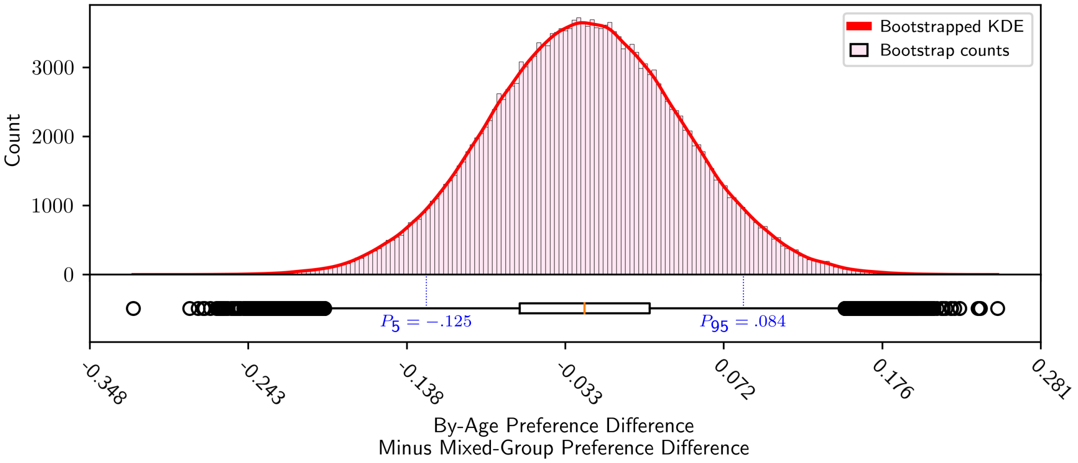
Dimension 45: child-related/toy-related

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.668875$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

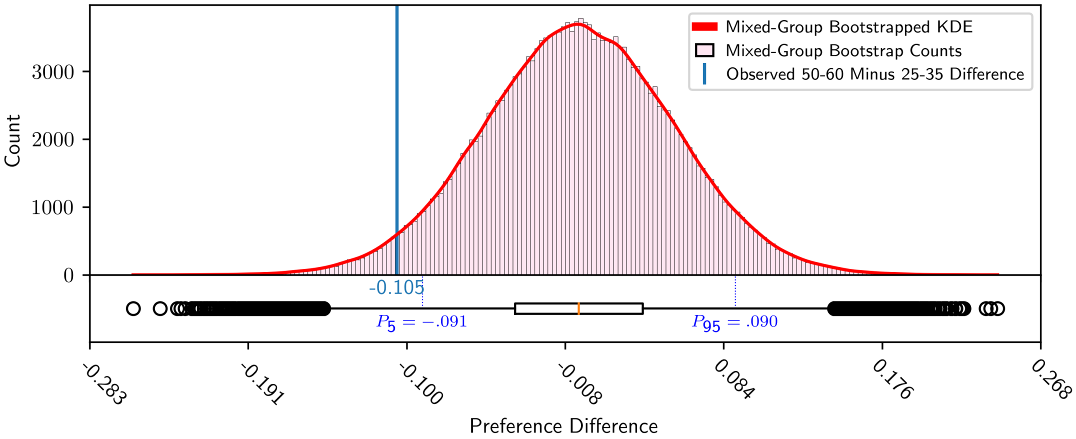
Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



$\mathrm{CI}_{90} = [-0.125, 0.084]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

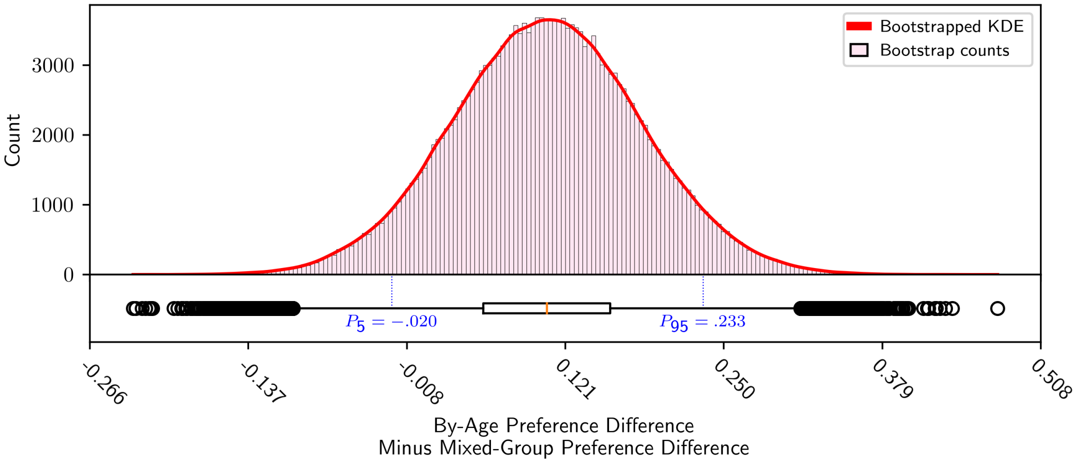
Dimension 46: medicine-related

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.056115$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

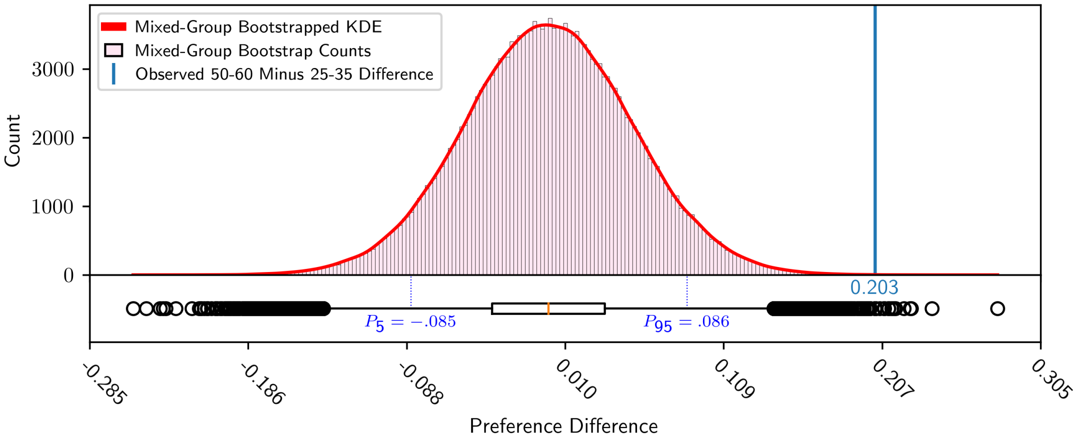
Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



$\mathrm{CI}_{90} = [-0.020, 0.233]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

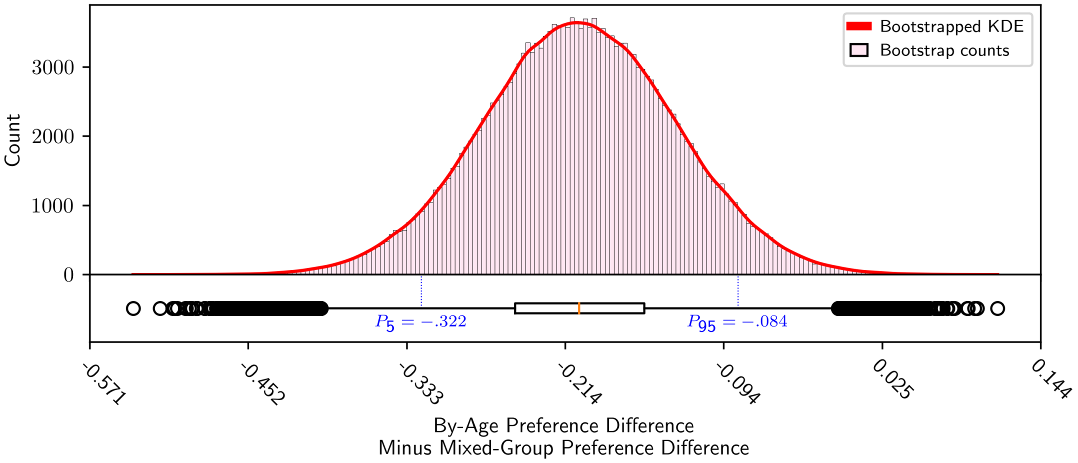
Dimension 47: has grating

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.00017$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

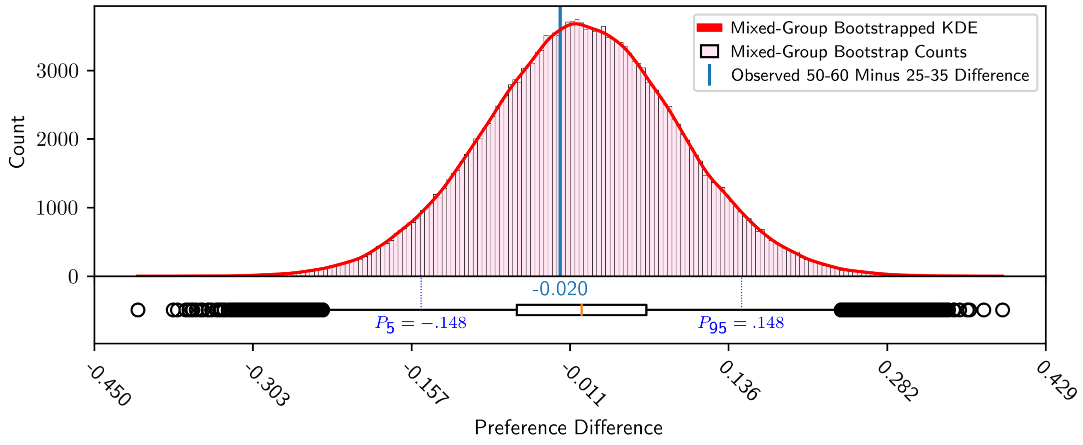
Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



$\mathrm{CI}_{90} = [-0.322, -0.084]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

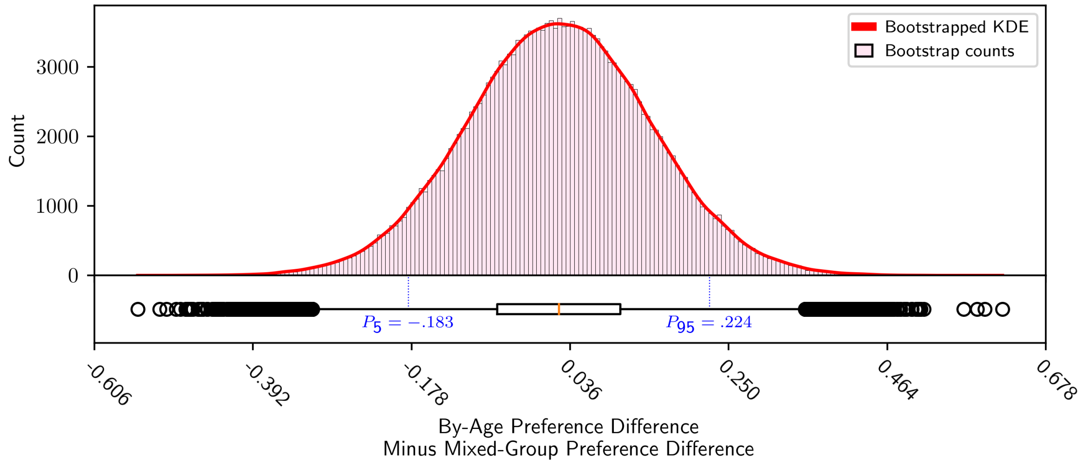
Dimension 48: handicraft-related

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.825045$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

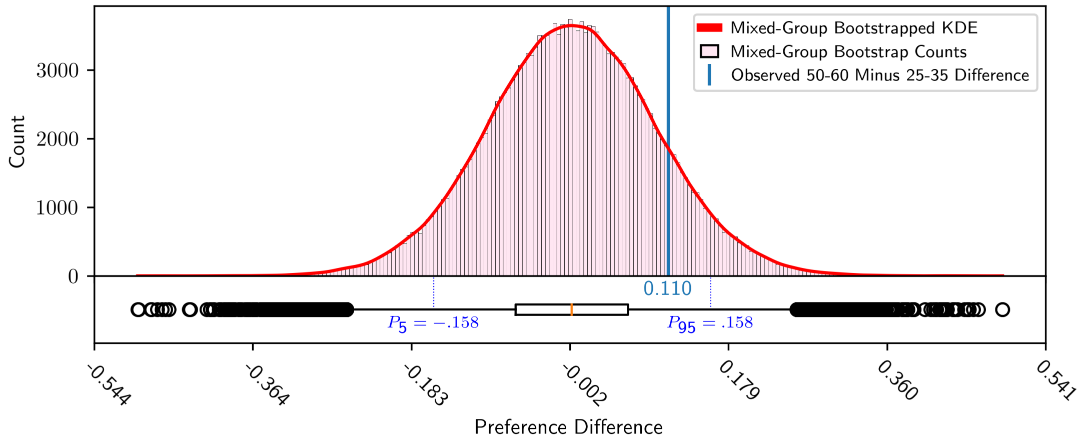
Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)



$\mathrm{CI}_{90} = [-0.183, 0.224]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

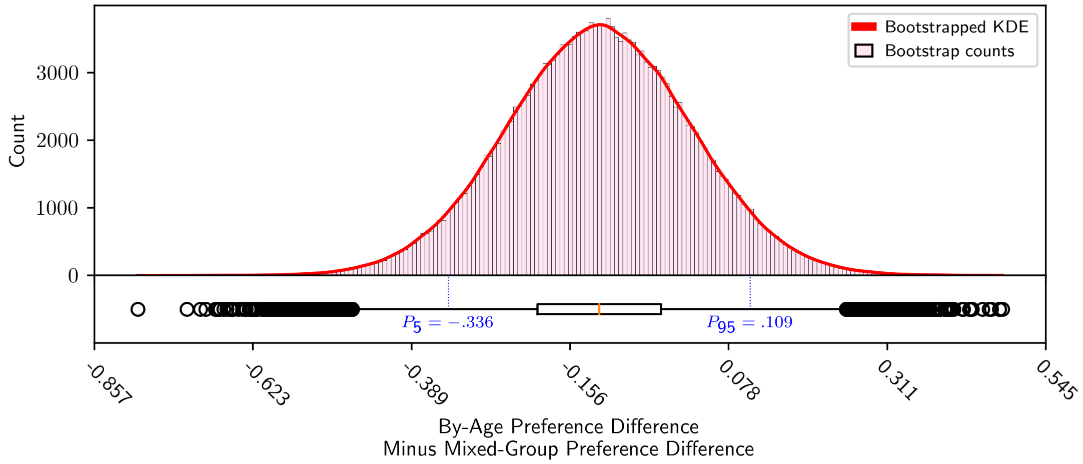
Dimension 49: cylindrical/conical

Observed By-Age Preference Difference (50-60 Minus 25-35)
vs. Bootstrapped Mixed Differences
(Bootstrap Count of 200000)



$p = 0.247015$
for obtaining the observed age-group difference
under the null hypothesis that their differences follow the mixed-group bootstrap distribution

Bootstrapped By-Age Preference Differences (Age 50-60 Minus 25-35)
Minus Bootstrapped Mixed-Group Differences
(Bootstrap Count of 200000)

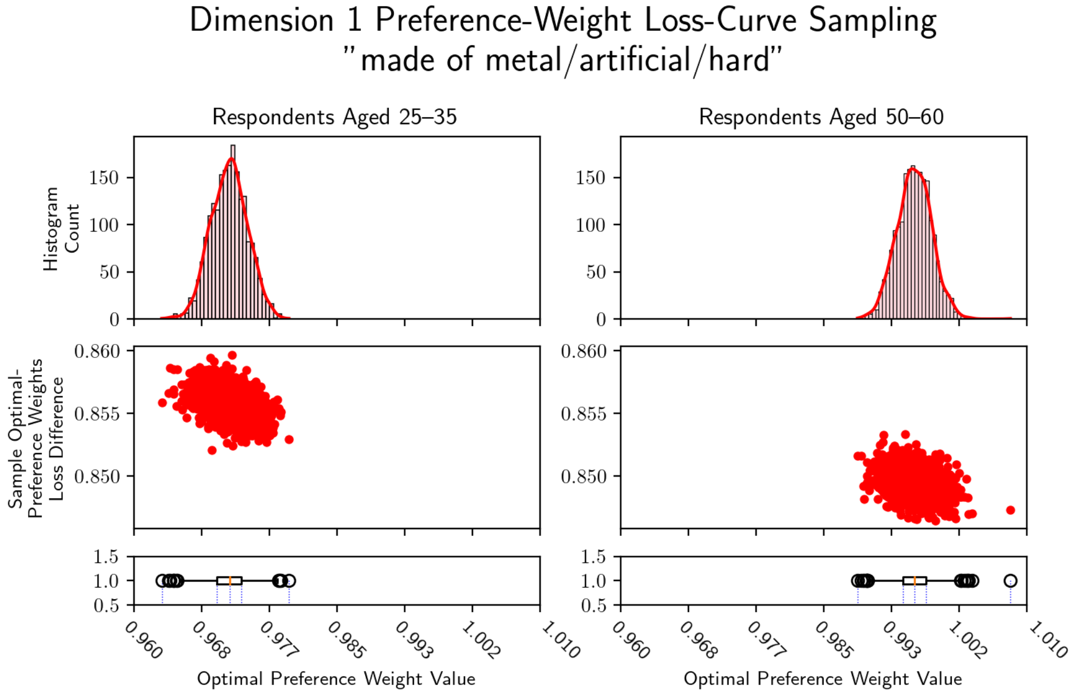


$\mathrm{CI}_{90} = [-0.336, 0.109]$
for the distribution of subtraction of the difference between random groups from the difference between ages
when bootstrapping mixed and separated 25-35- and 50-60-year-olds' responses
(zero's presence in CI_{90} indicates some sample age differences are exceeded by chance when bootstrapping)

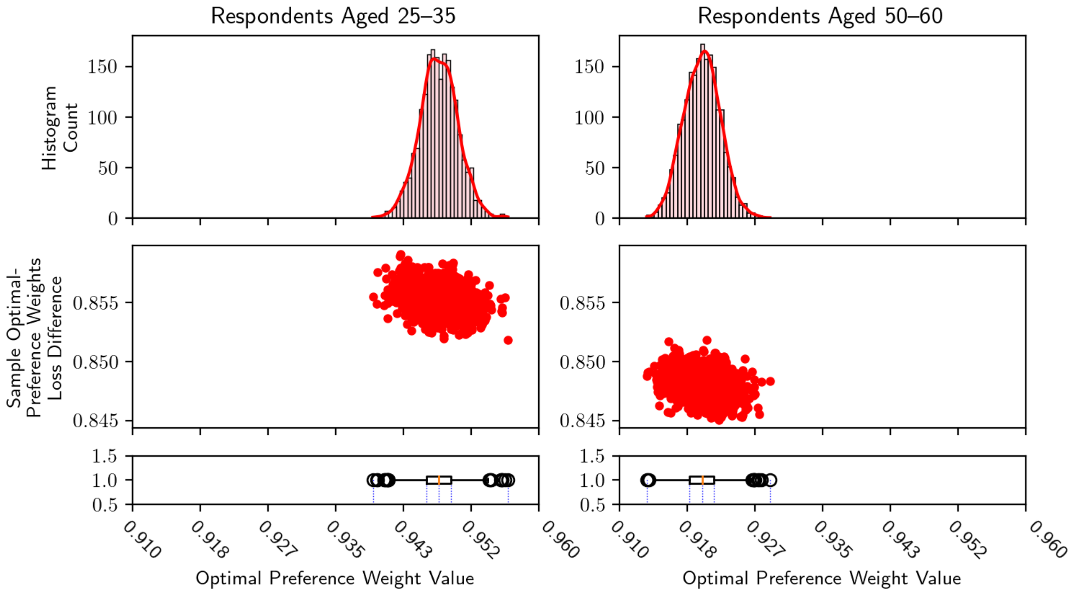
A.2 By-Age Preference-Weight Sampling Distributions

The distributions of bootstrapped likeness-preference weights for each age group. The bootstrapped preferences for the 25–35-year-olds' responses are on the left, while 50–60-year-olds' are on the right. The x-axis gives found preference-weight values, and both distributions are on the same horizontal scale.

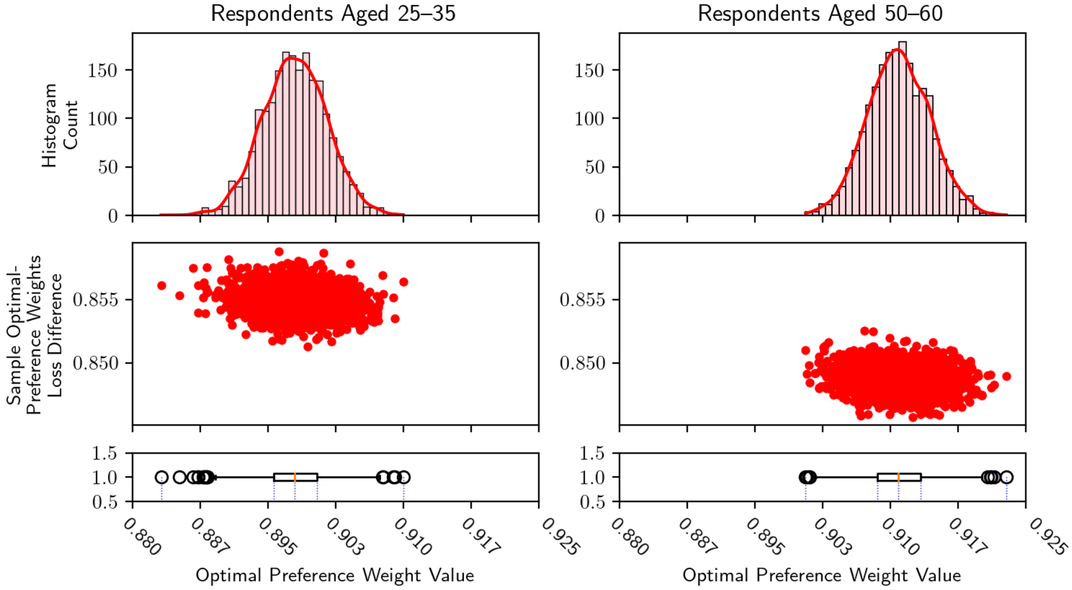
Lack of overlap between the two age groups' preference-weight distributions is correlated with significant differences between each age groups' preferences, but these distributions should not be directly used for statistical testing due to some level of difference being explained by stochasticity. See Section A.1 for distributions valid for statistical significance and Section 3.4.8.2 for an explanation of those distributions.



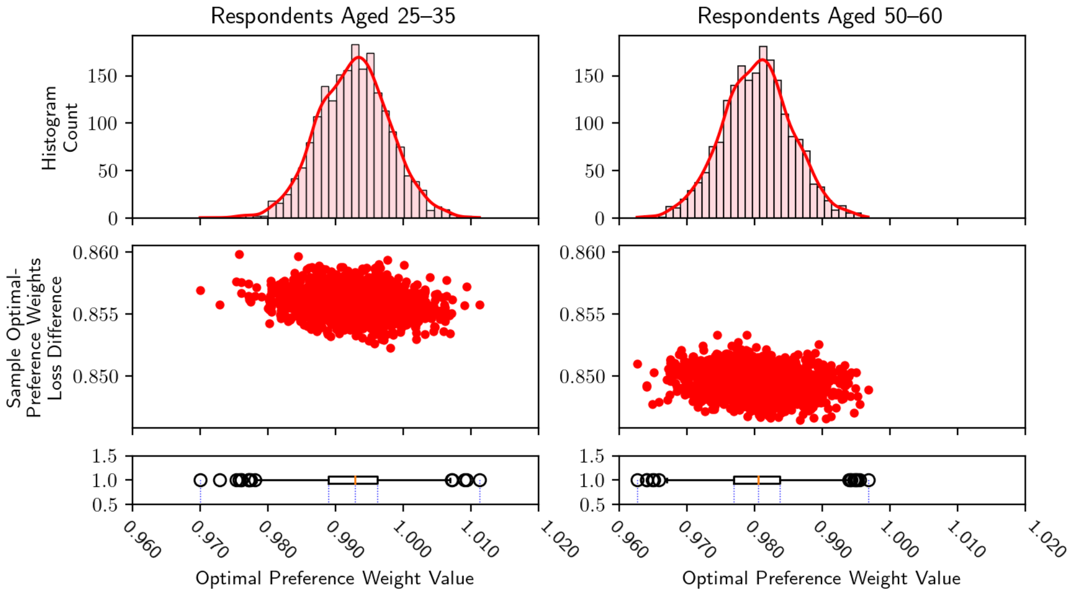
Dimension 2 Preference-Weight Loss-Curve Sampling
"food-related/eating-related/kitchen-related"



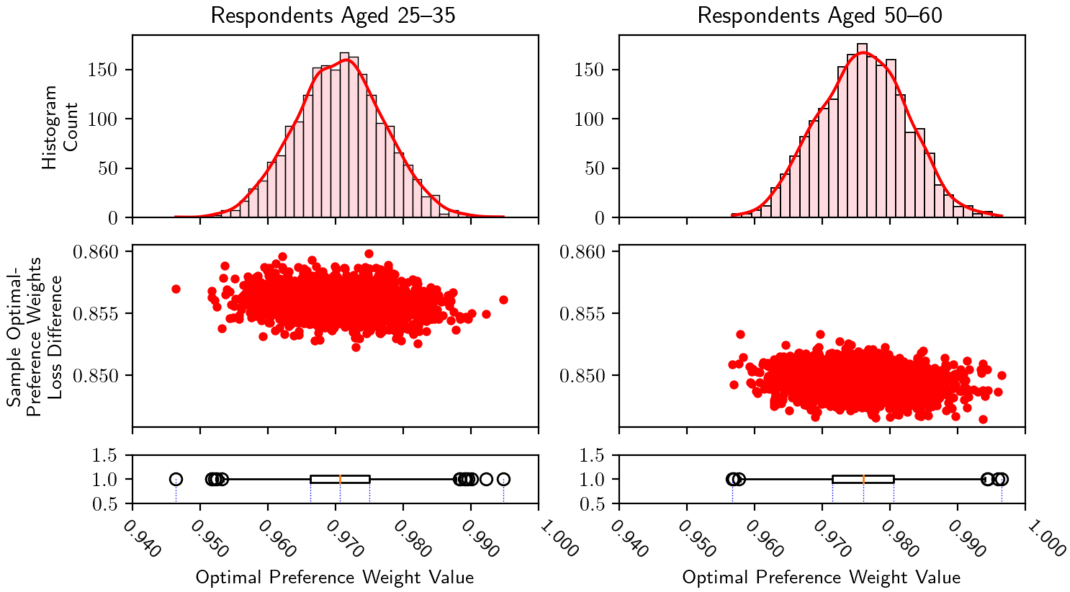
Dimension 3 Preference-Weight Loss-Curve Sampling
"animal-related/organic"



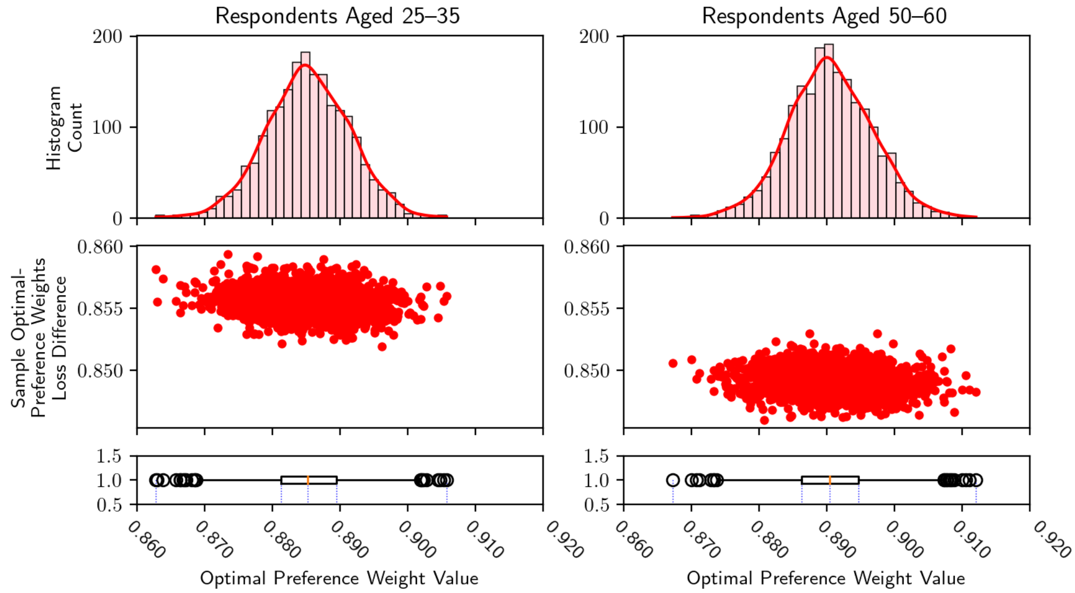
Dimension 4 Preference-Weight Loss-Curve Sampling
"clothing-related/fabric/covering"



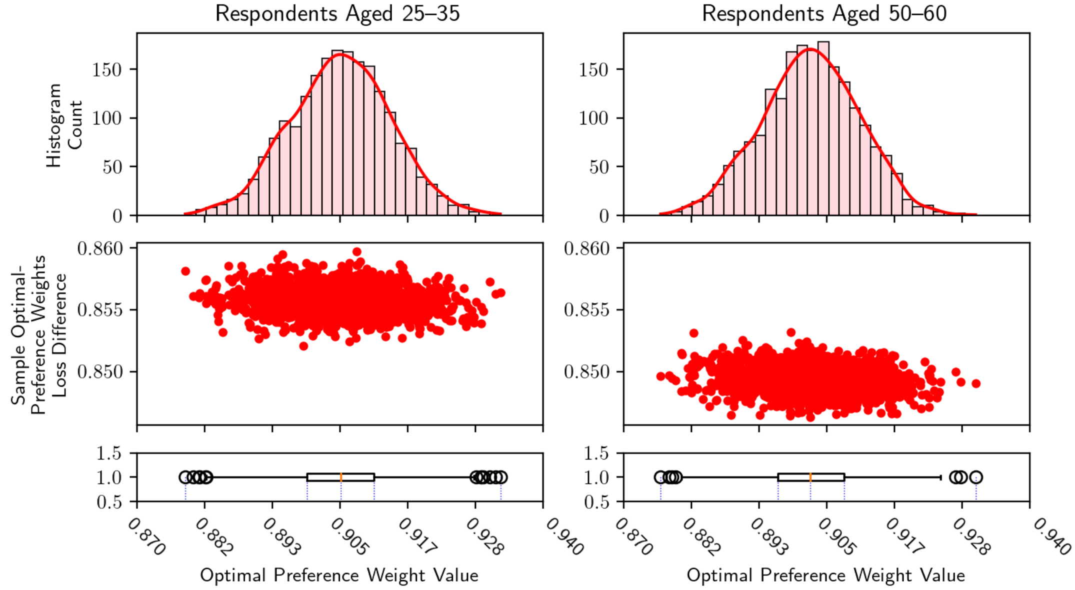
Dimension 5 Preference-Weight Loss-Curve Sampling
"furniture-related/household-related/artifact"

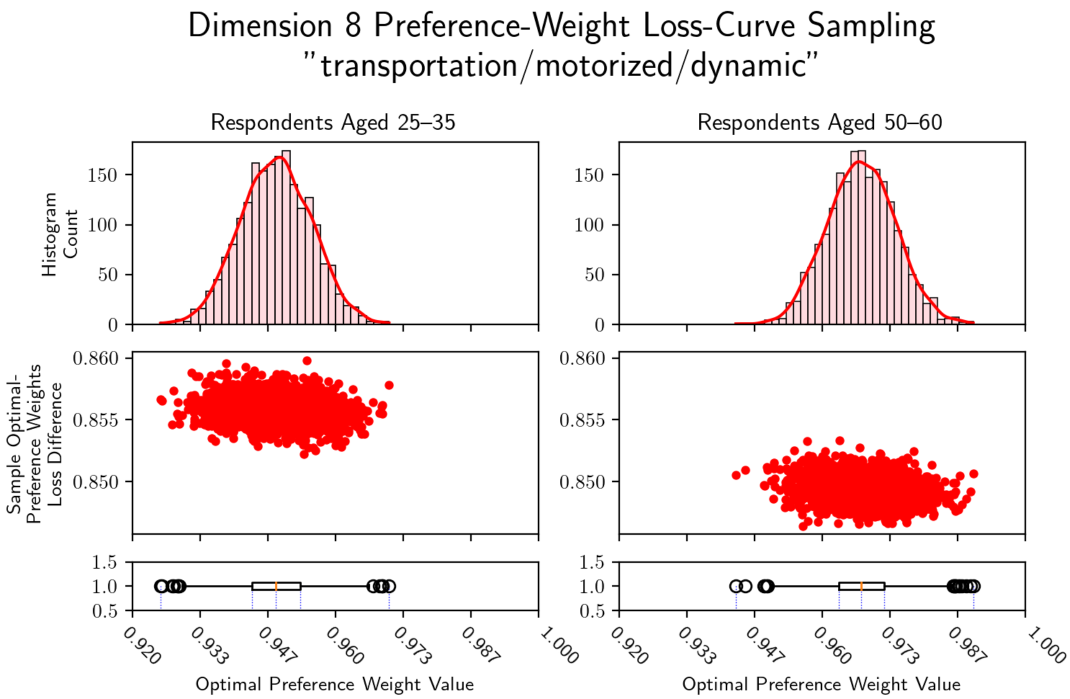


Dimension 6 Preference-Weight Loss-Curve Sampling
"plant-related/green"



Dimension 7 Preference-Weight Loss-Curve Sampling
"outdoors-related"

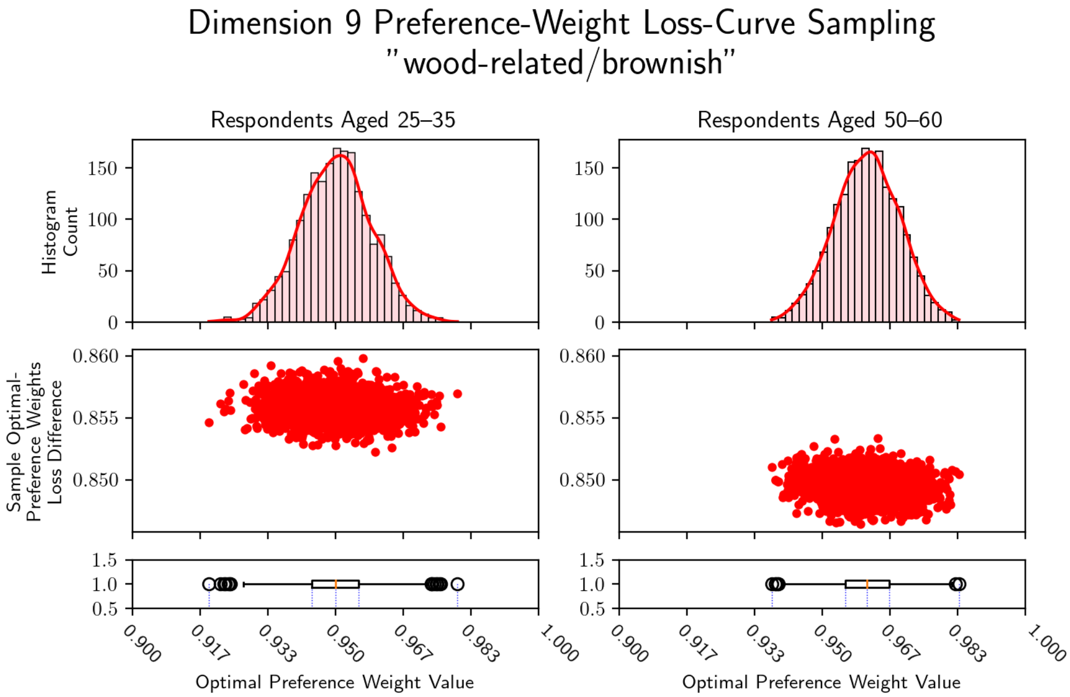




Dimension 8 Preference-Weight Loss-Curve Sampling
"transportation/motorized/dynamic"

Respondents Aged 25–35

Respondents Aged 50–60

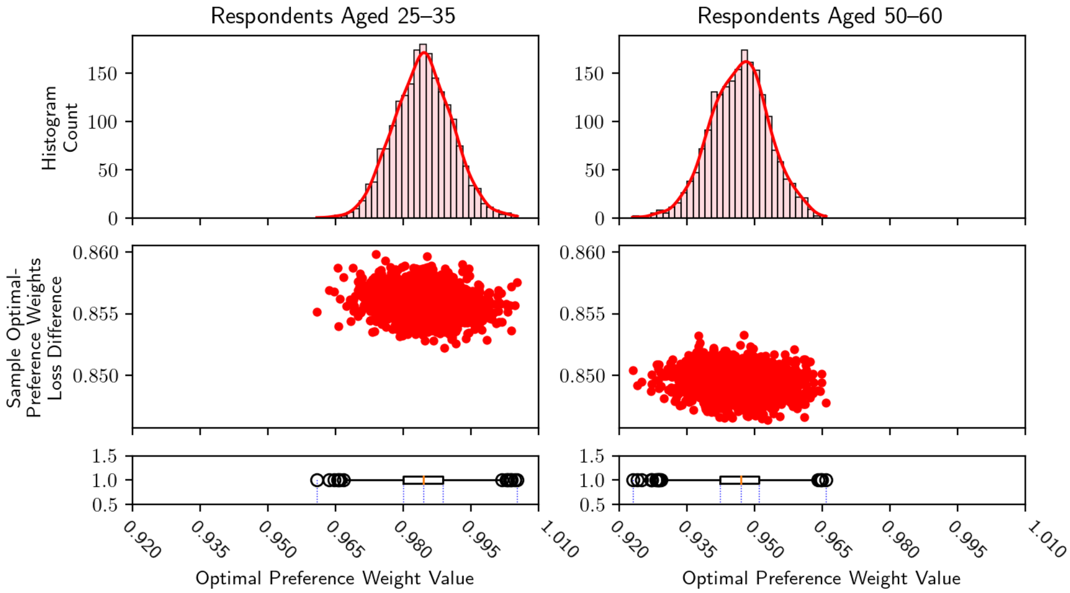


Dimension 9 Preference-Weight Loss-Curve Sampling
"wood-related/brownish"

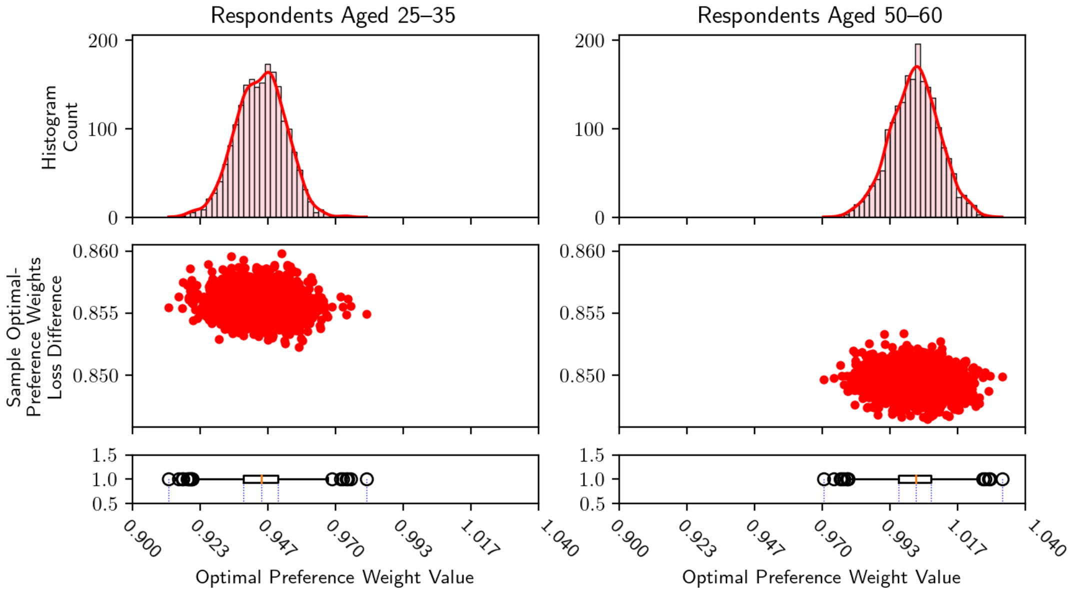
Respondents Aged 25–35

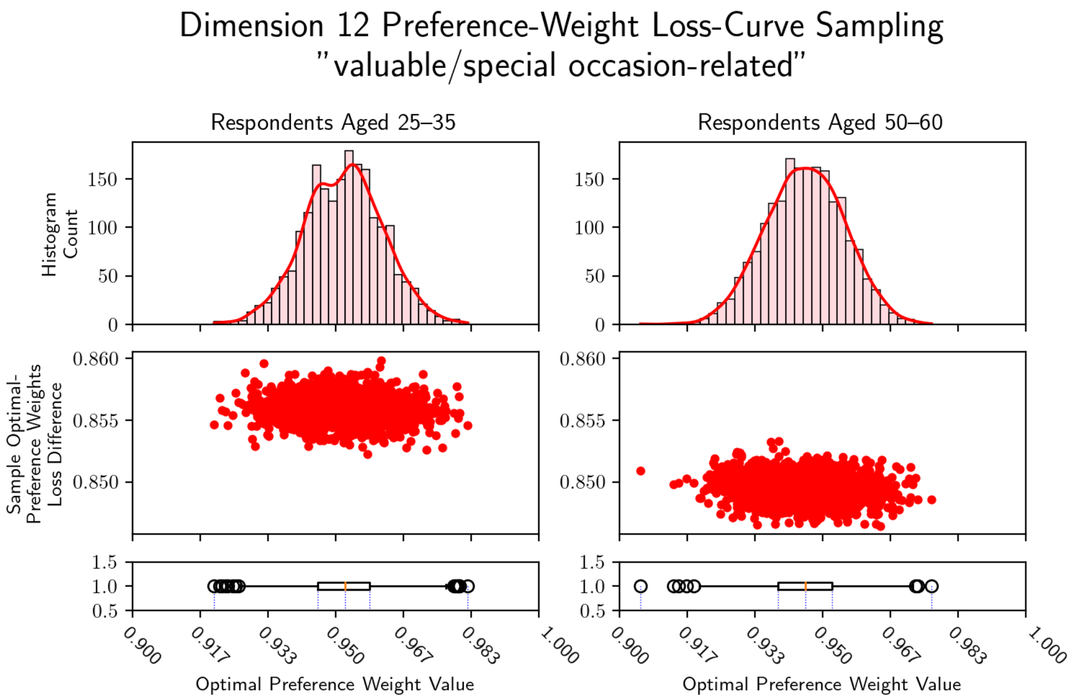
Respondents Aged 50–60

Dimension 10 Preference-Weight Loss-Curve Sampling
"body part-related"



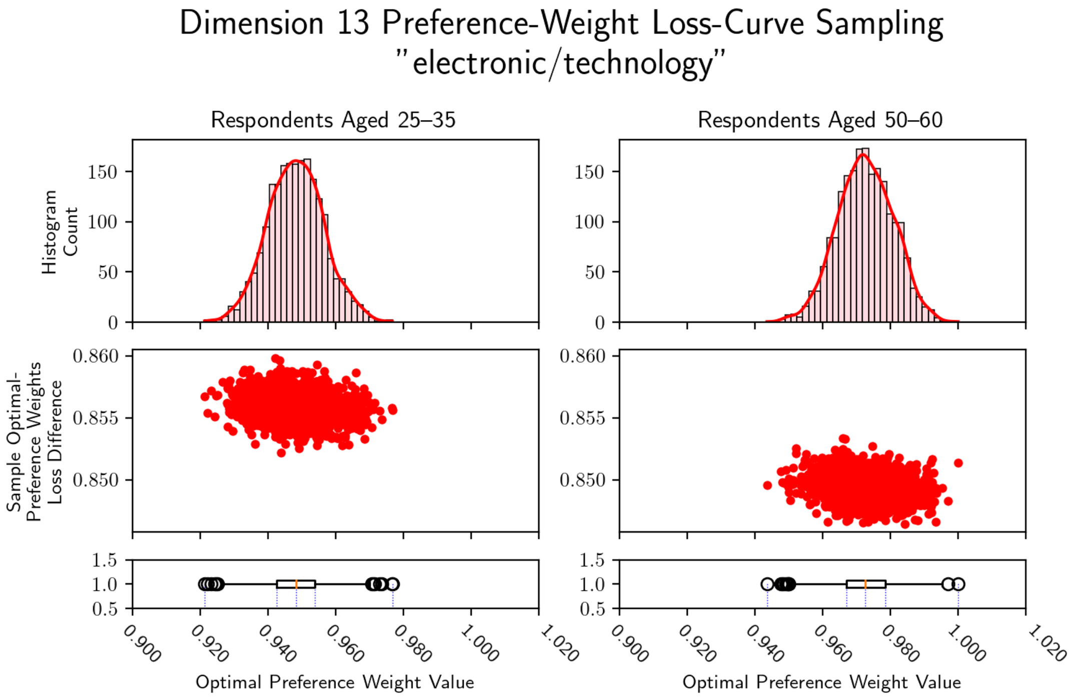
Dimension 11 Preference-Weight Loss-Curve Sampling
"colorful"




Dimension 12 Preference-Weight Loss-Curve Sampling
"valuable/special occasion-related"

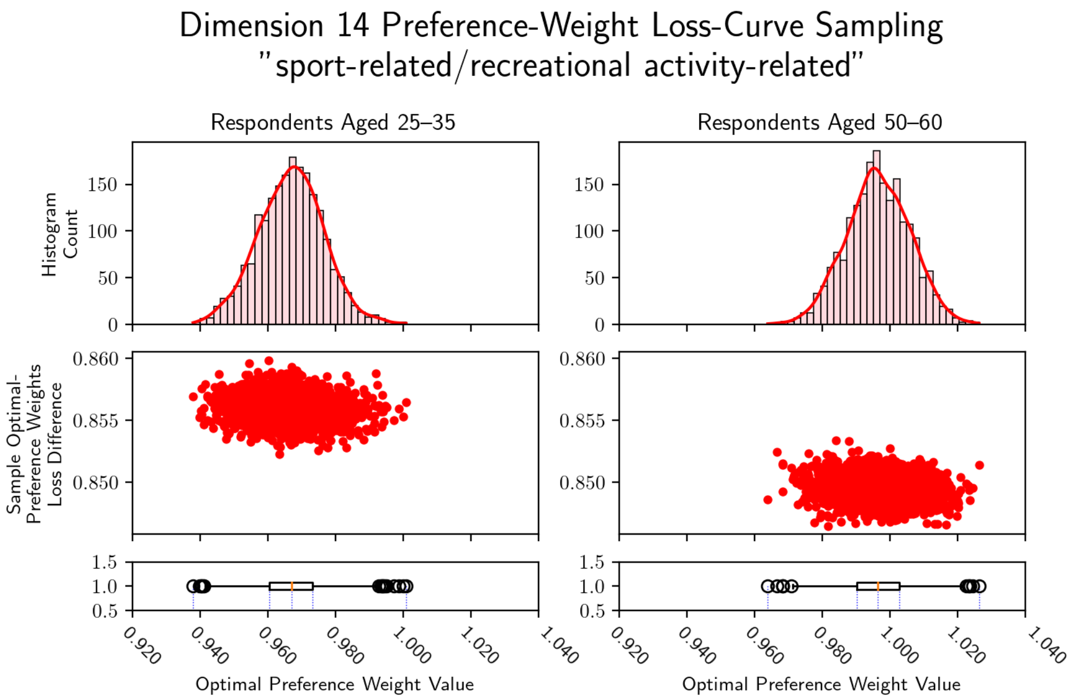
Respondents Aged 25–35

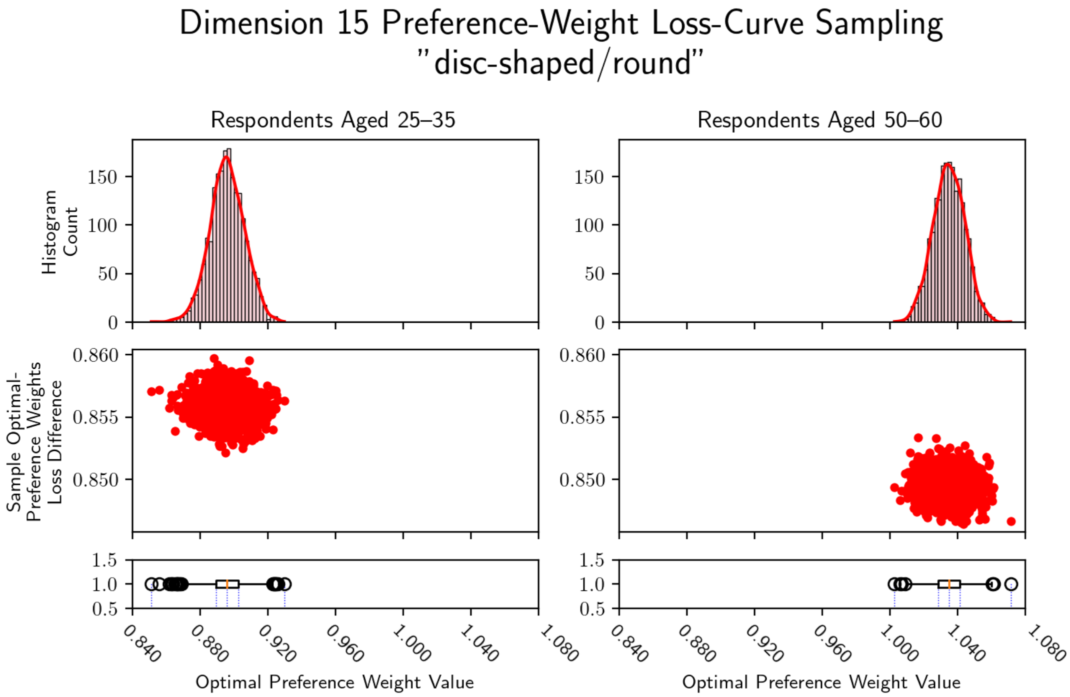
Respondents Aged 50–60


Dimension 13 Preference-Weight Loss-Curve Sampling
"electronic/technology"

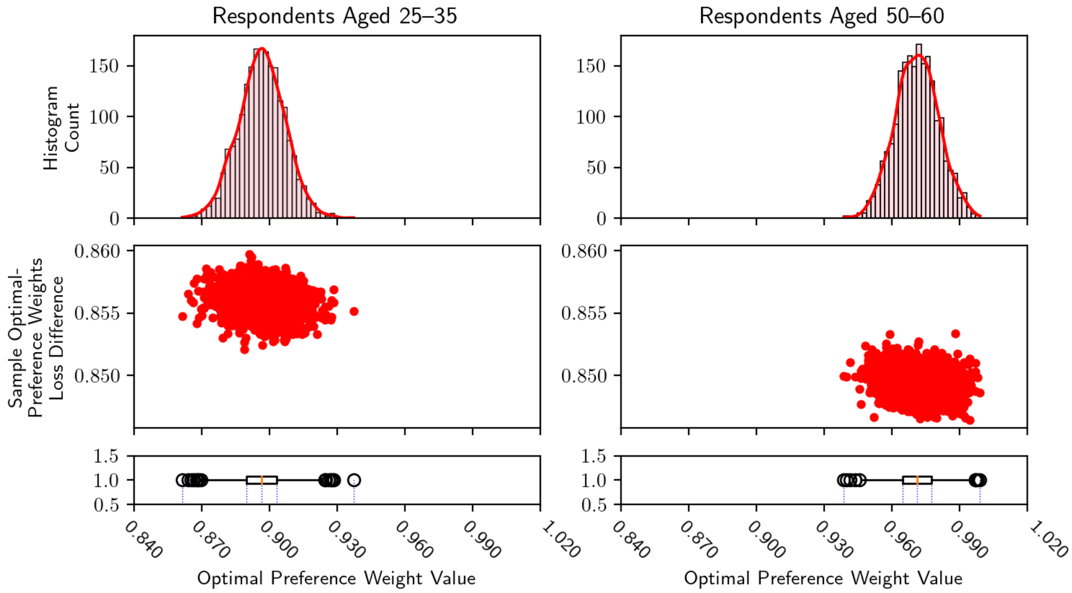
Respondents Aged 25–35

Respondents Aged 50–60

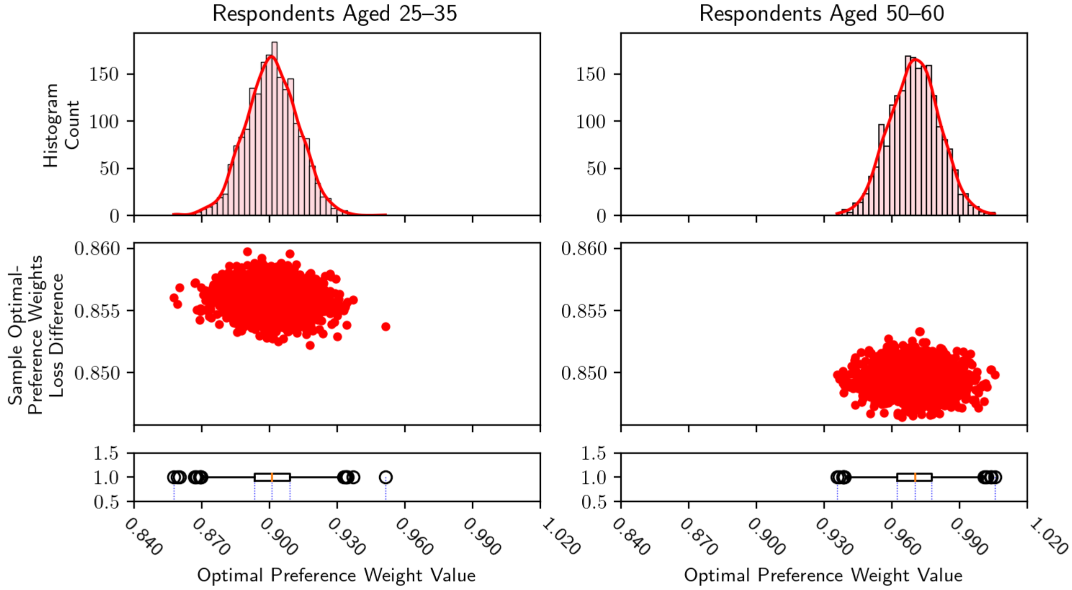

Dimension 14 Preference-Weight Loss-Curve Sampling
"sport-related/recreational activity-related"
Respondents Aged 25–35
Respondents Aged 50–60
Histogram Count
0
50
100
150
Sample Optimal-Preference Weights Loss Difference
0.850
0.855
0.860
0.5
1.0
1.5
0.920
0.940
0.960
0.980
1.000
1.020
1.040
Optimal Preference Weight Value


Dimension 15 Preference-Weight Loss-Curve Sampling
"disc-shaped/round"
Respondents Aged 25–35
Respondents Aged 50–60
Histogram Count
0
50
100
150
Sample Optimal-Preference Weights Loss Difference
0.850
0.855
0.860
0.5
1.0
1.5
0.840
0.880
0.920
0.960
1.000
1.040
1.080
Optimal Preference Weight Value

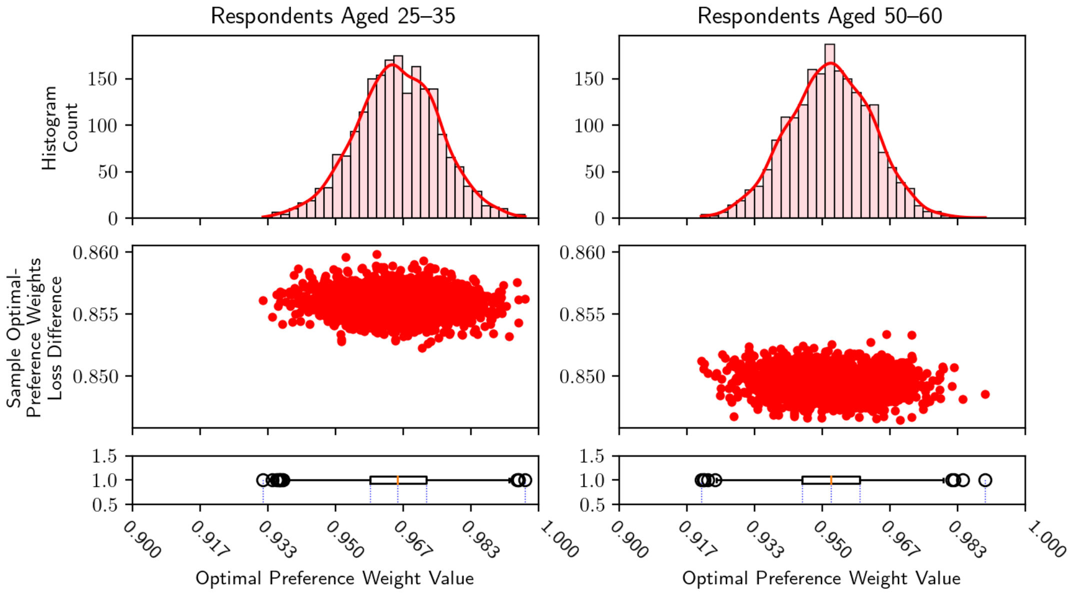
Dimension 16 Preference-Weight Loss-Curve Sampling
"tool-related"



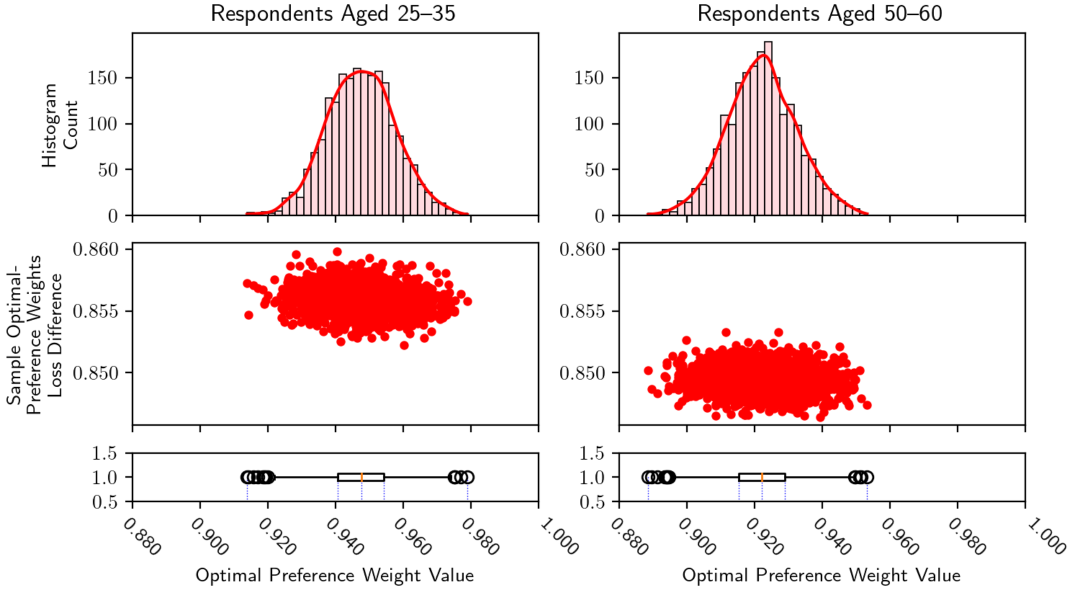
Dimension 17 Preference-Weight Loss-Curve Sampling
"many small things/course pattern"



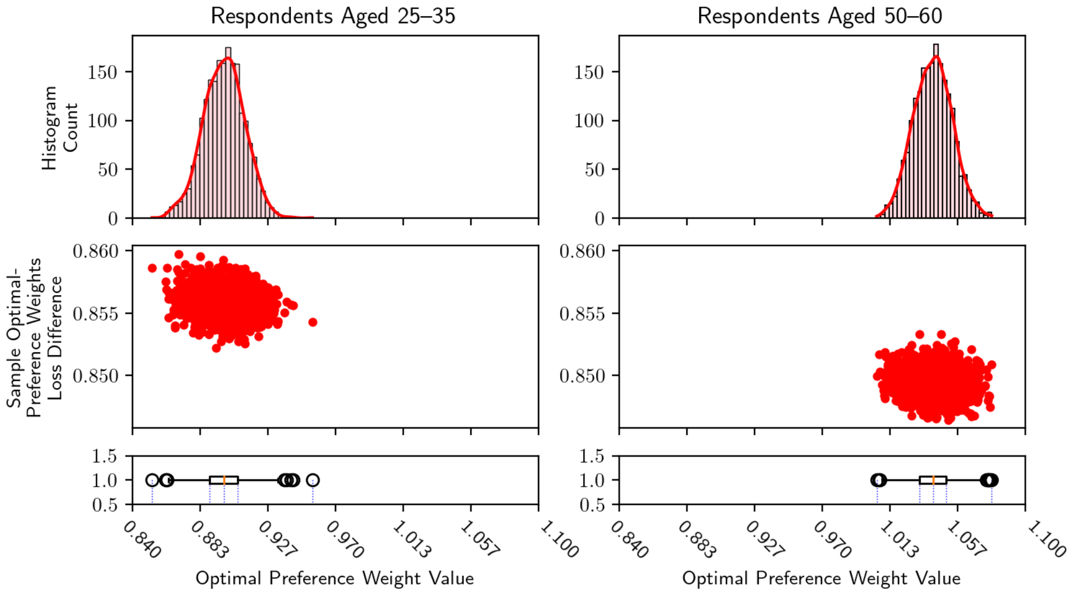
Dimension 18 Preference-Weight Loss-Curve Sampling
"paper-related/thin/flat/text-related"



Dimension 19 Preference-Weight Loss-Curve Sampling
"fluid-related/drink-related"

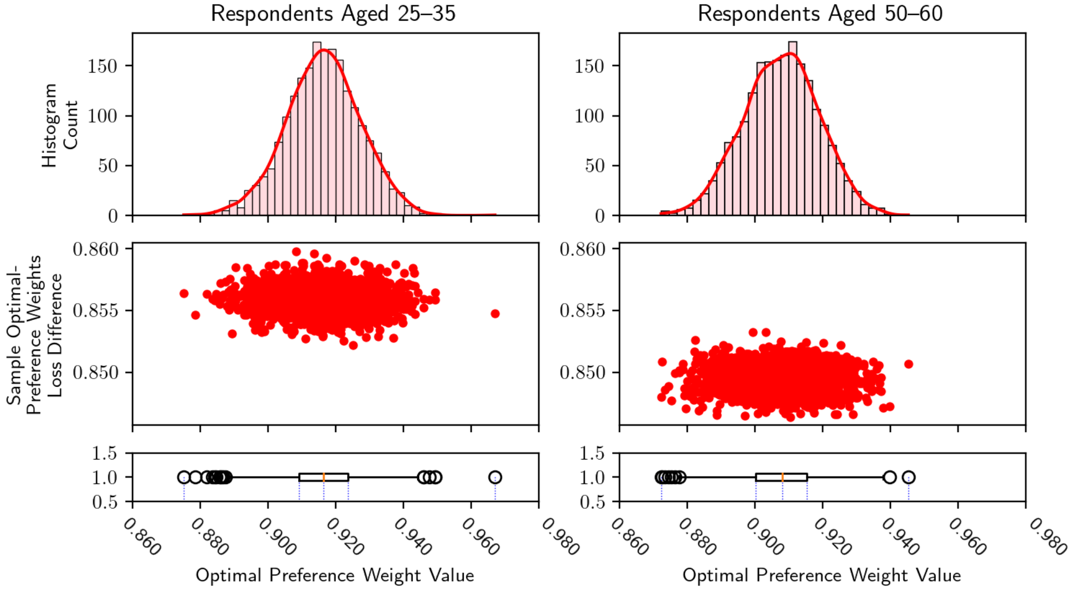


Dimension 20 Preference-Weight Loss-Curve Sampling
"long/thin"


Respondents Aged 25–35

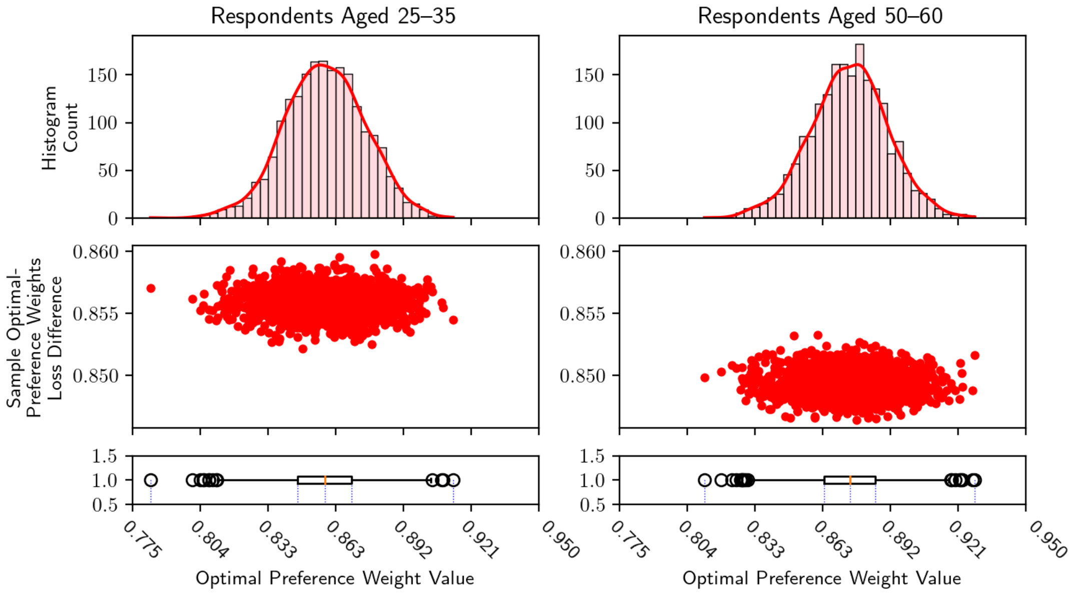
Respondents Aged 50–60

Dimension 21 Preference-Weight Loss-Curve Sampling
"water-related/blue"

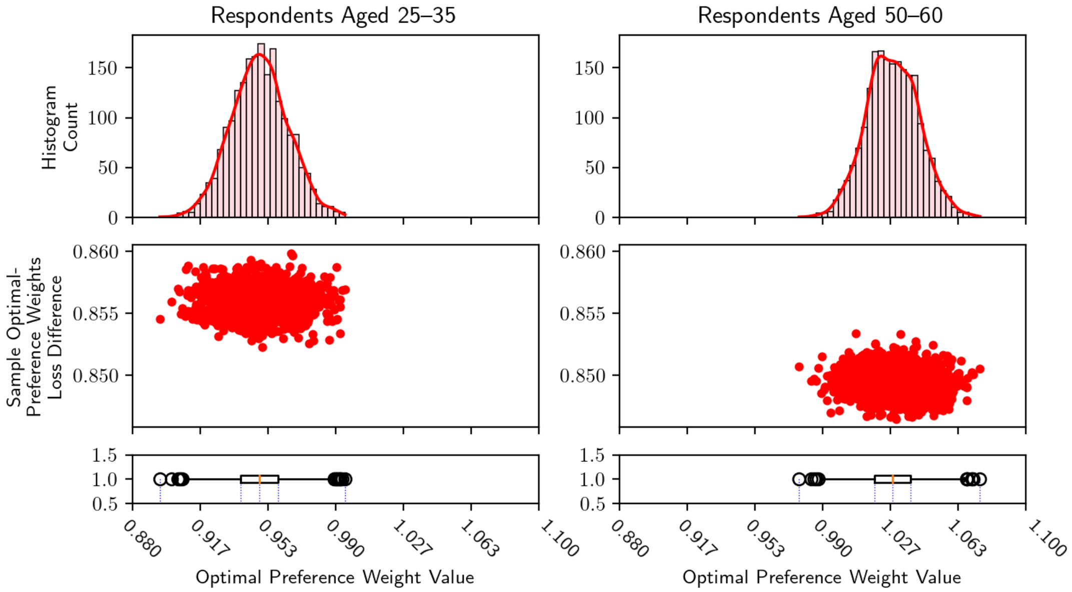

Respondents Aged 25–35

Respondents Aged 50–60

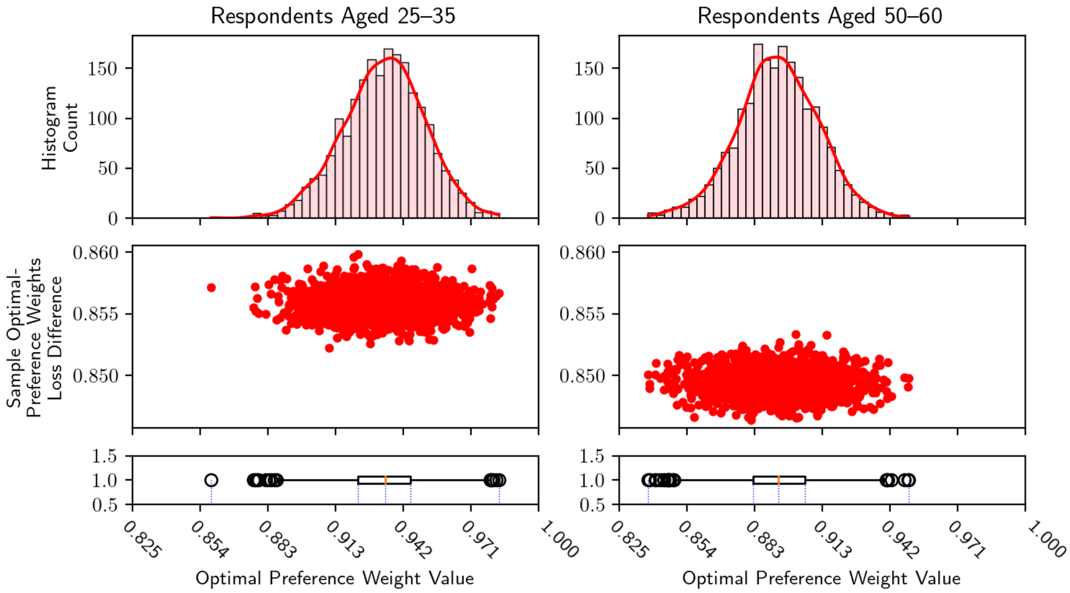
Dimension 22 Preference-Weight Loss-Curve Sampling
"powdery/fine-scale pattern"



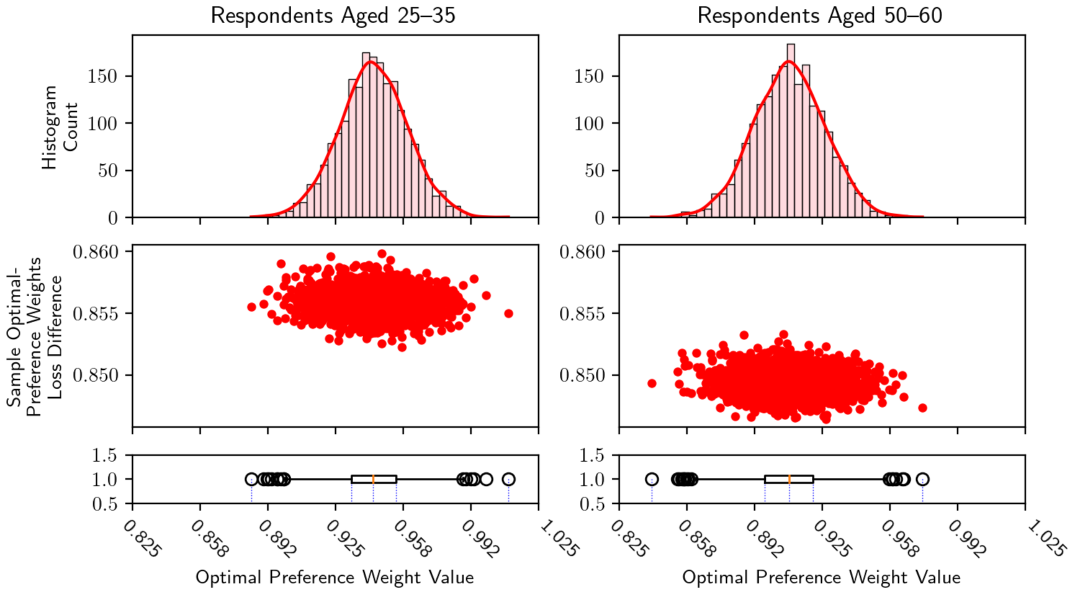
Dimension 23 Preference-Weight Loss-Curve Sampling
"red"



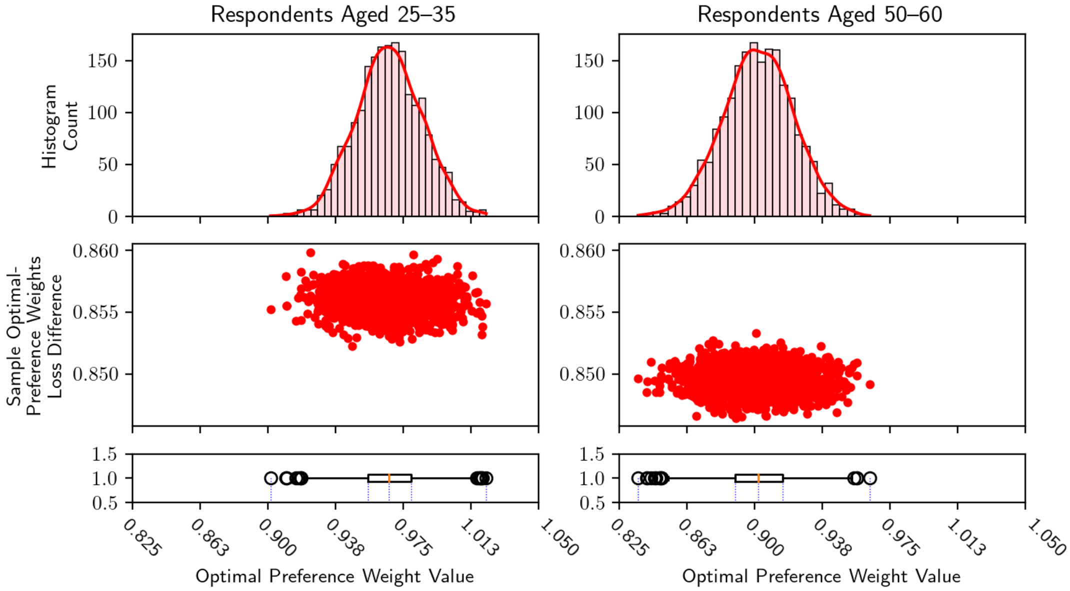
Dimension 24 Preference-Weight Loss-Curve Sampling
"feminine (stereotypically)/decorative"



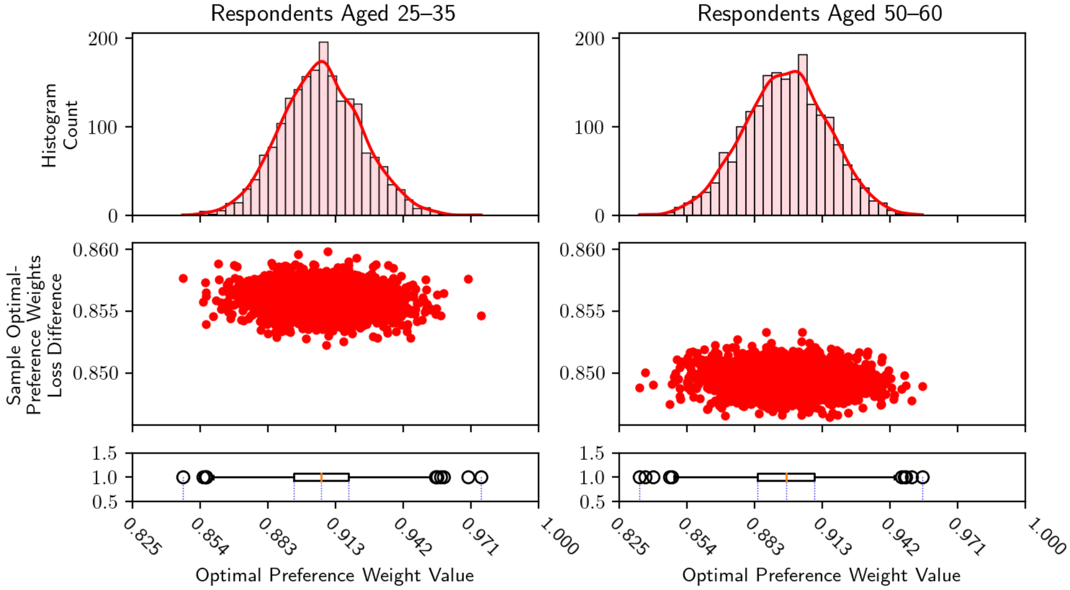
Dimension 25 Preference-Weight Loss-Curve Sampling
"bathroom-related/sanitary"



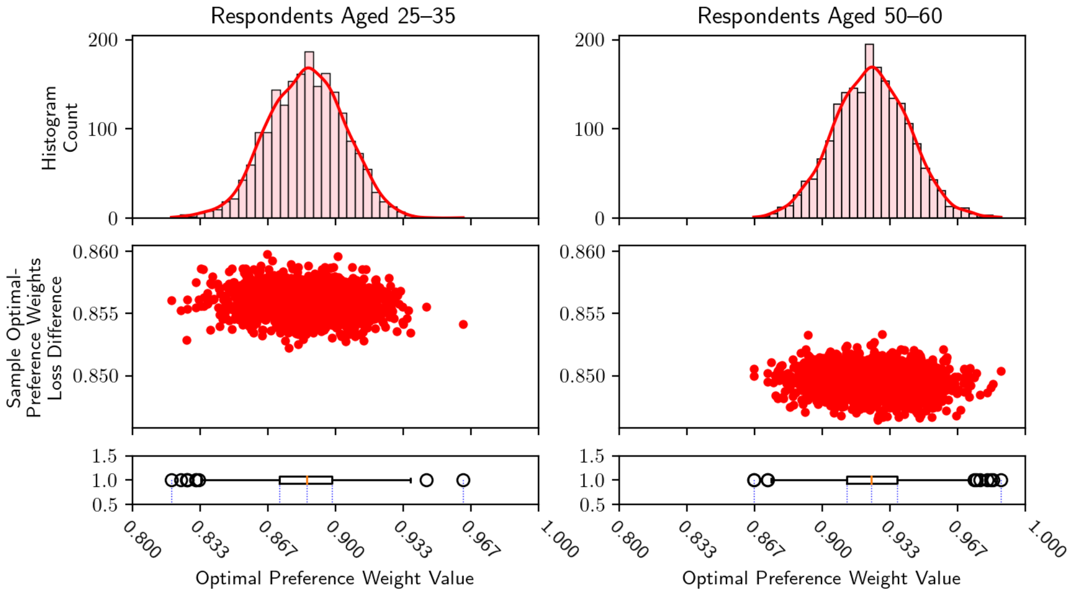
Dimension 26 Preference-Weight Loss-Curve Sampling
"black/noble"



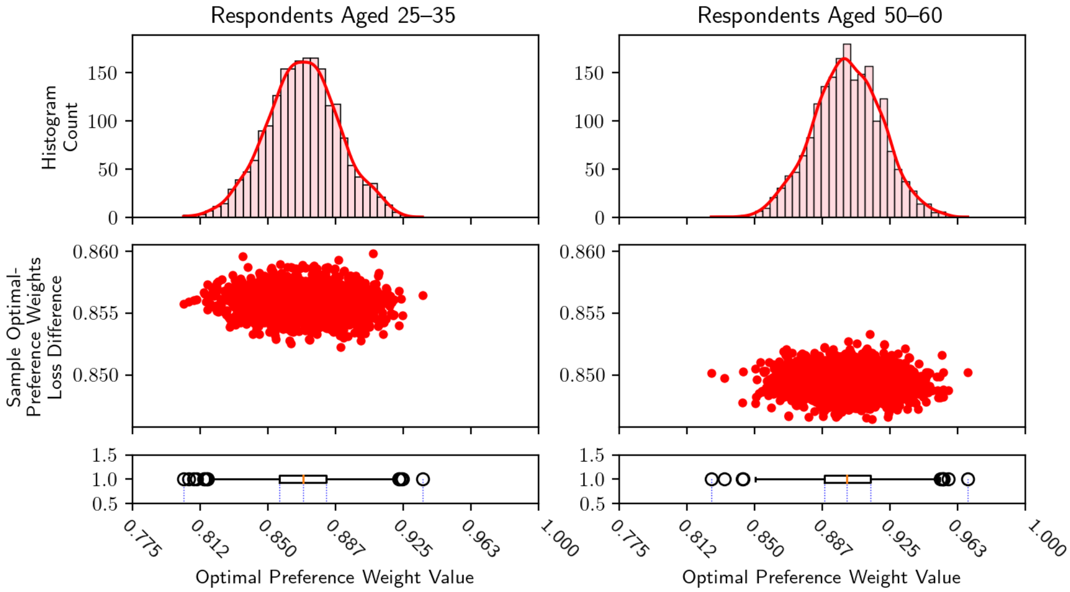
Dimension 27 Preference-Weight Loss-Curve Sampling
"weapon/danger-related/violence"

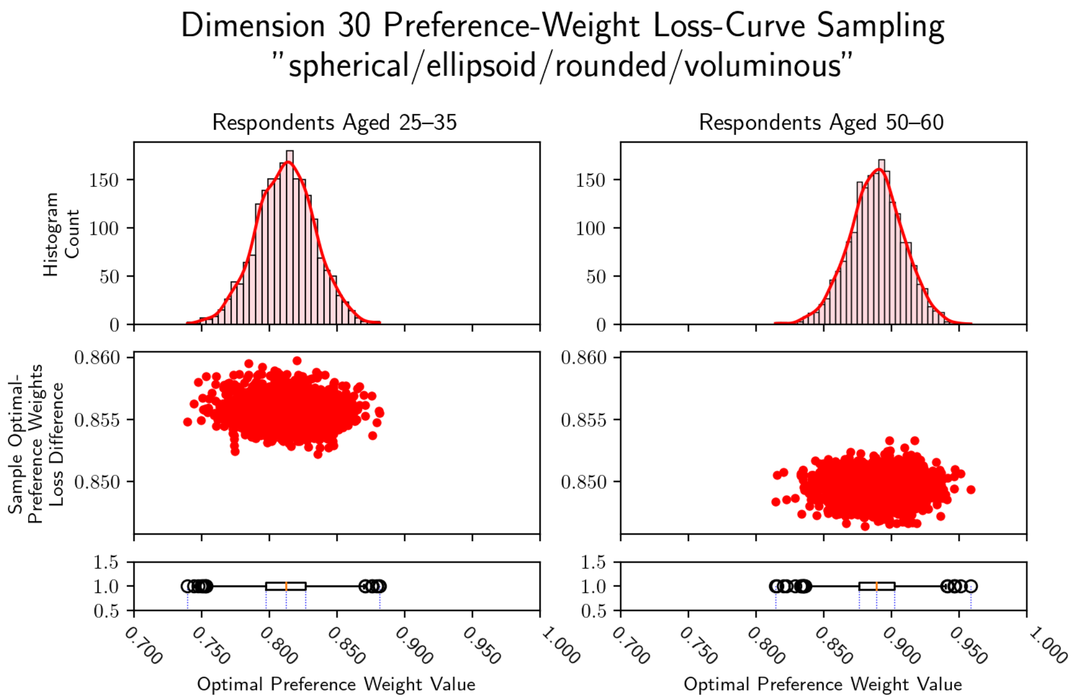


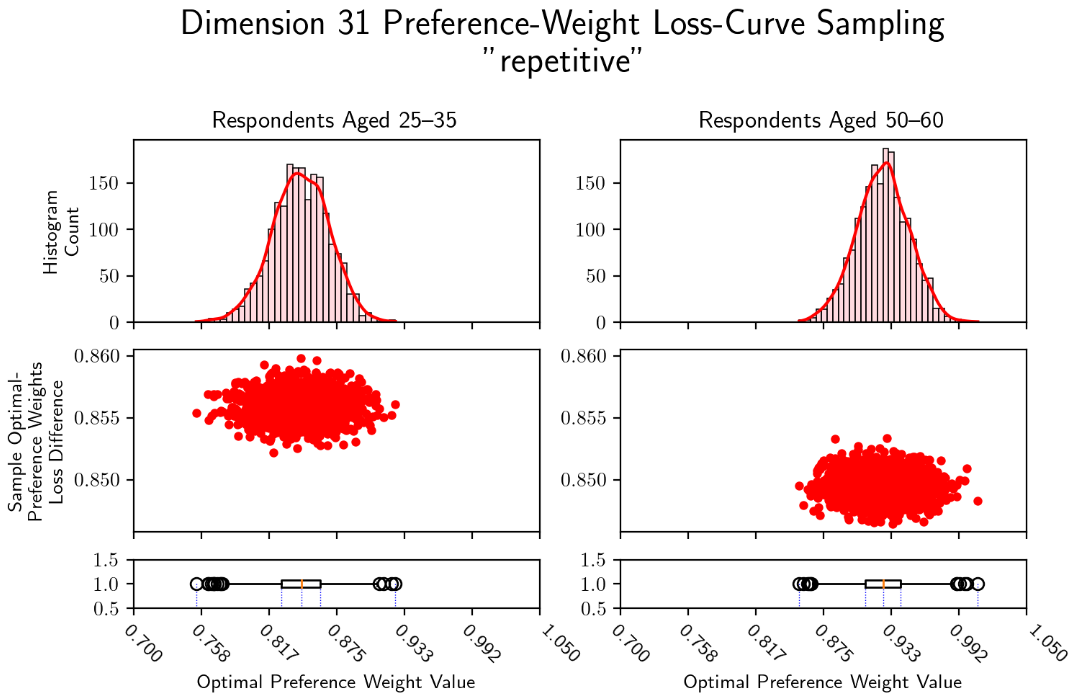
Dimension 28 Preference-Weight Loss-Curve Sampling
"musical instrument-related/noise-related"



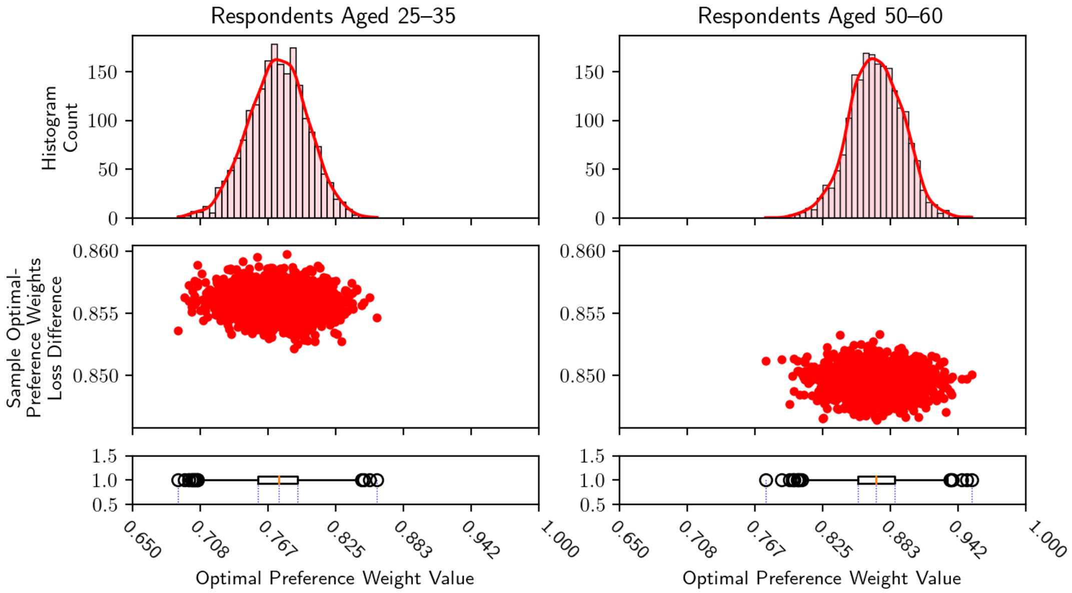
Dimension 29 Preference-Weight Loss-Curve Sampling
"sky-related/flying-related/floating-related"



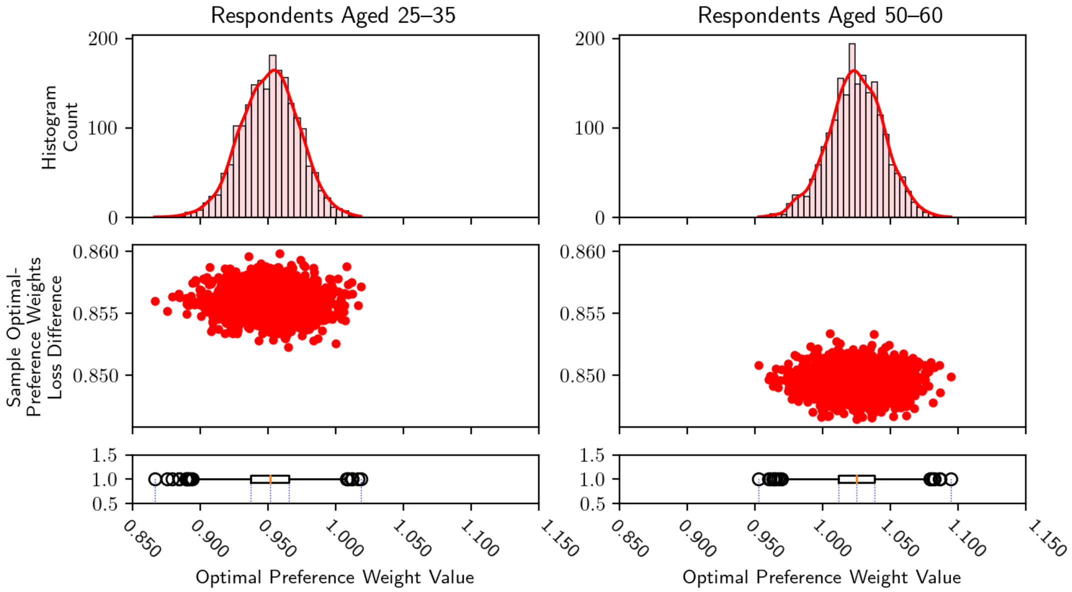

Dimension 30 Preference-Weight Loss-Curve Sampling
"spherical/ellipsoid/rounded/voluminous"
Respondents Aged 25–35
Respondents Aged 50–60
Histogram Count
150
100
50
0
Sample Optimal-Preference Weights Loss Difference
0.860
0.855
0.850
1.5
1.0
0.5
0.700
0.750
0.800
0.850
0.900
0.950
1.000
Optimal Preference Weight Value


Dimension 31 Preference-Weight Loss-Curve Sampling
"repetitive"
Respondents Aged 25–35
Respondents Aged 50–60
Histogram Count
150
100
50
0
Sample Optimal-Preference Weights Loss Difference
0.860
0.855
0.850
1.5
1.0
0.5
0.700
0.758
0.817
0.875
0.933
0.992
1.050
Optimal Preference Weight Value

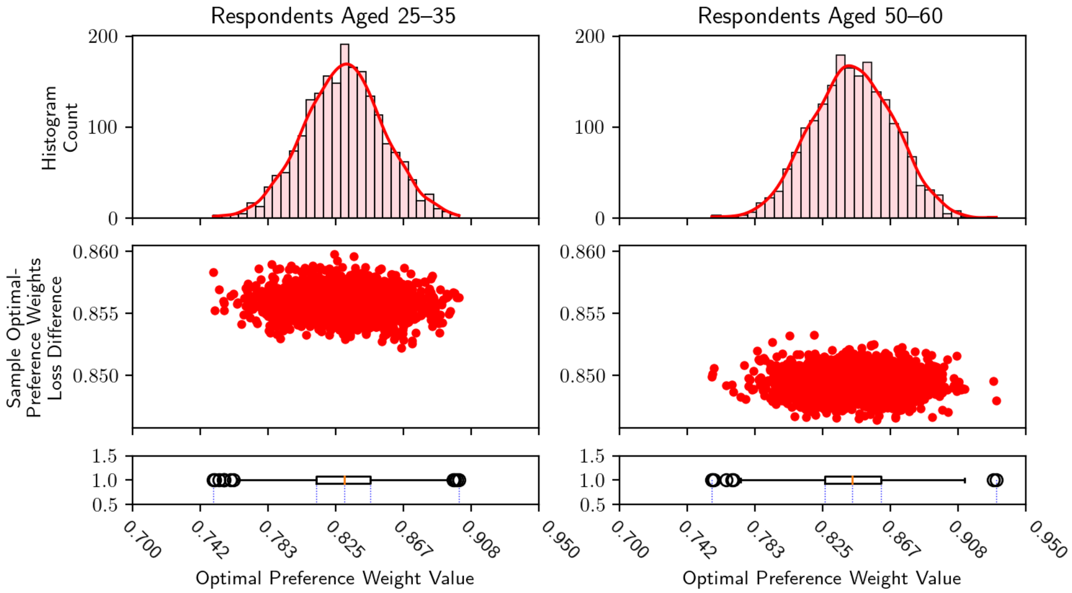
Dimension 32 Preference-Weight Loss-Curve Sampling
"flat/patterned"



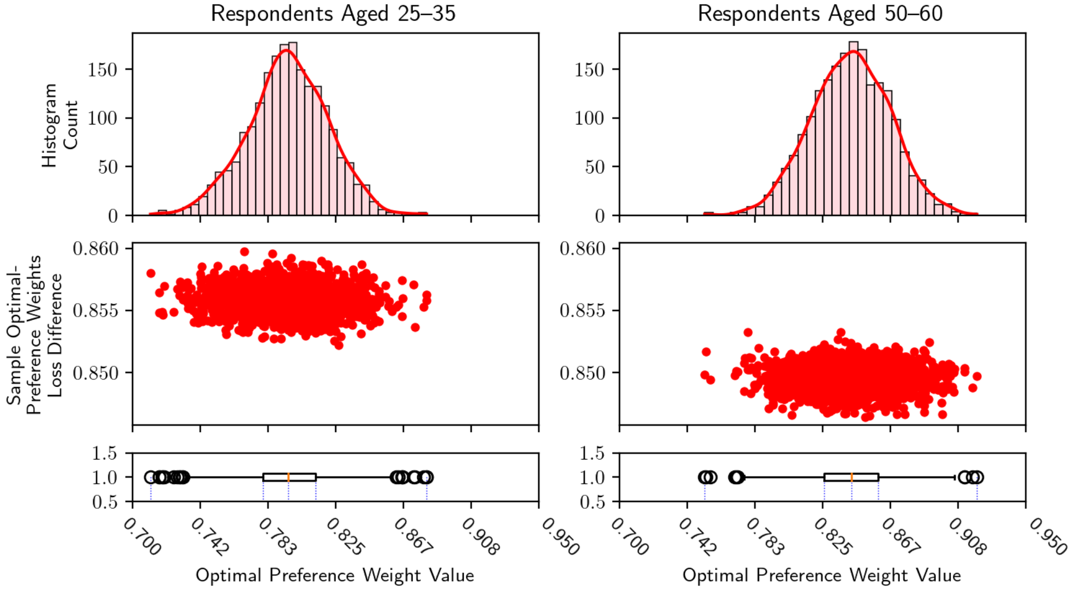
Dimension 33 Preference-Weight Loss-Curve Sampling
"white"



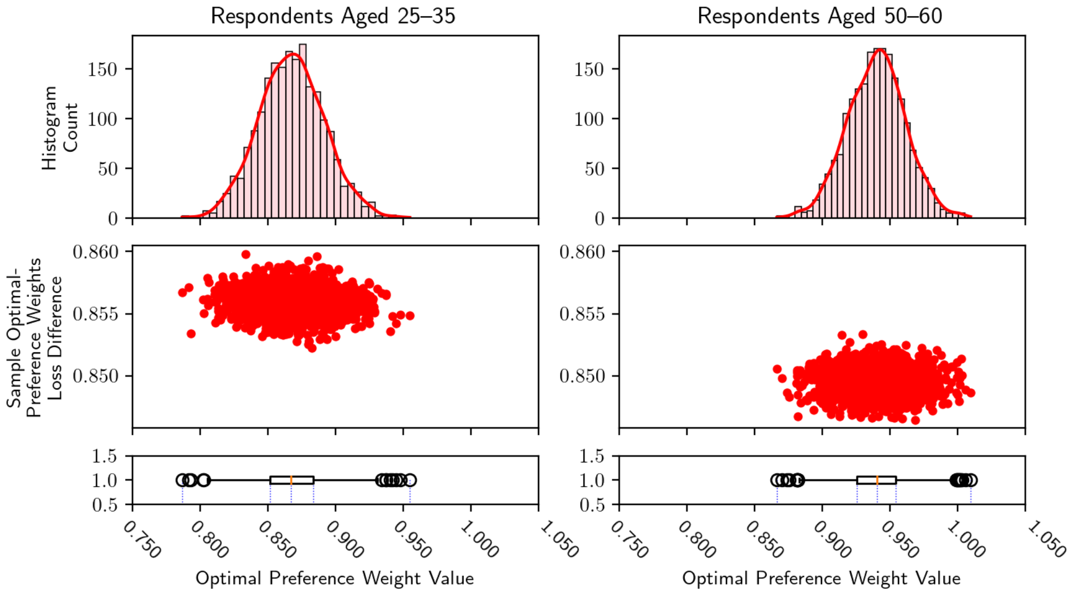
Dimension 34 Preference-Weight Loss-Curve Sampling
"thin/flat"



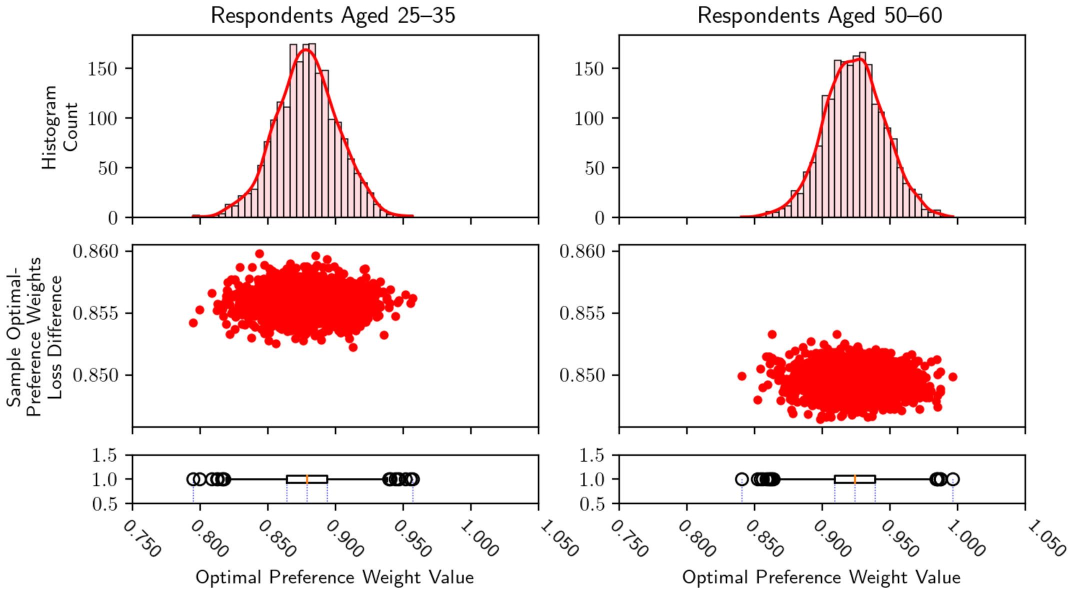
Dimension 35 Preference-Weight Loss-Curve Sampling
"disgusting/bugs"



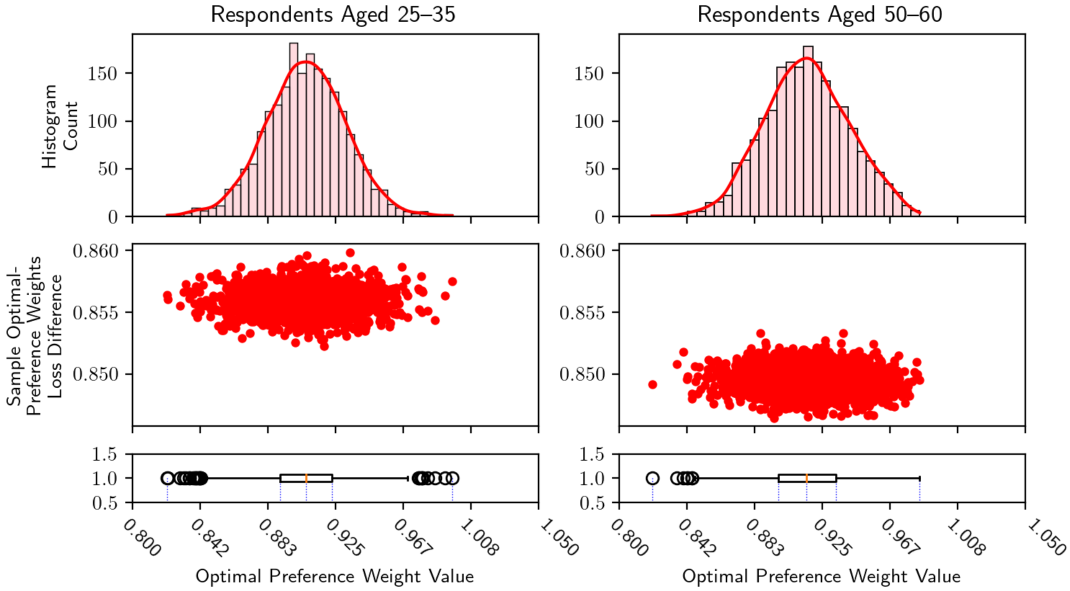
Dimension 36 Preference-Weight Loss-Curve Sampling
"string-related"



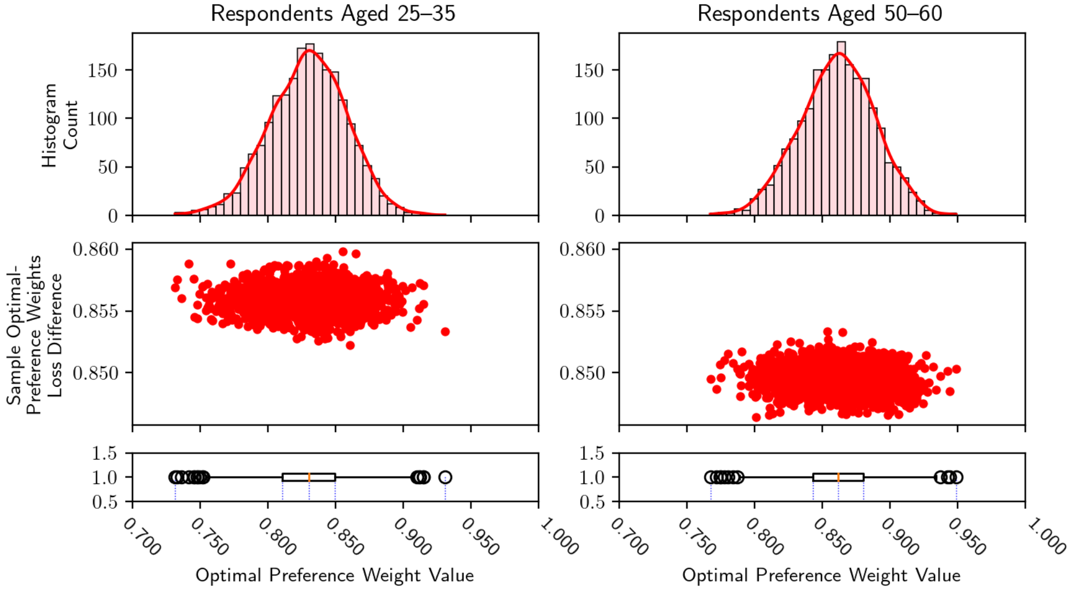
Dimension 37 Preference-Weight Loss-Curve Sampling
"arms/legs/skin-related"



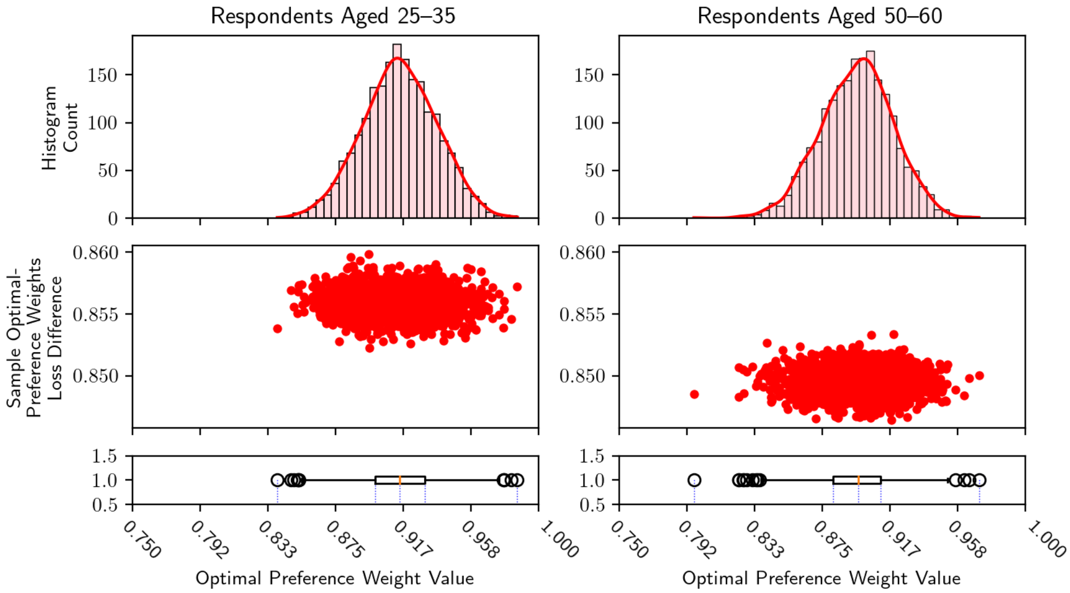
Dimension 38 Preference-Weight Loss-Curve Sampling
"shiny/transparent"



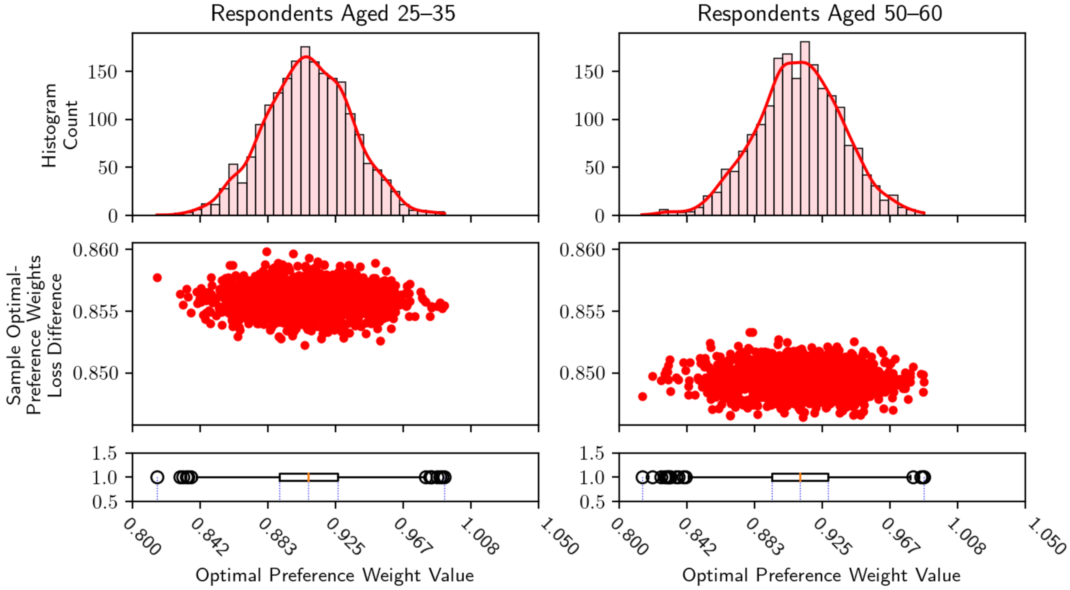
Dimension 39 Preference-Weight Loss-Curve Sampling
"construction-related/physical work-related"



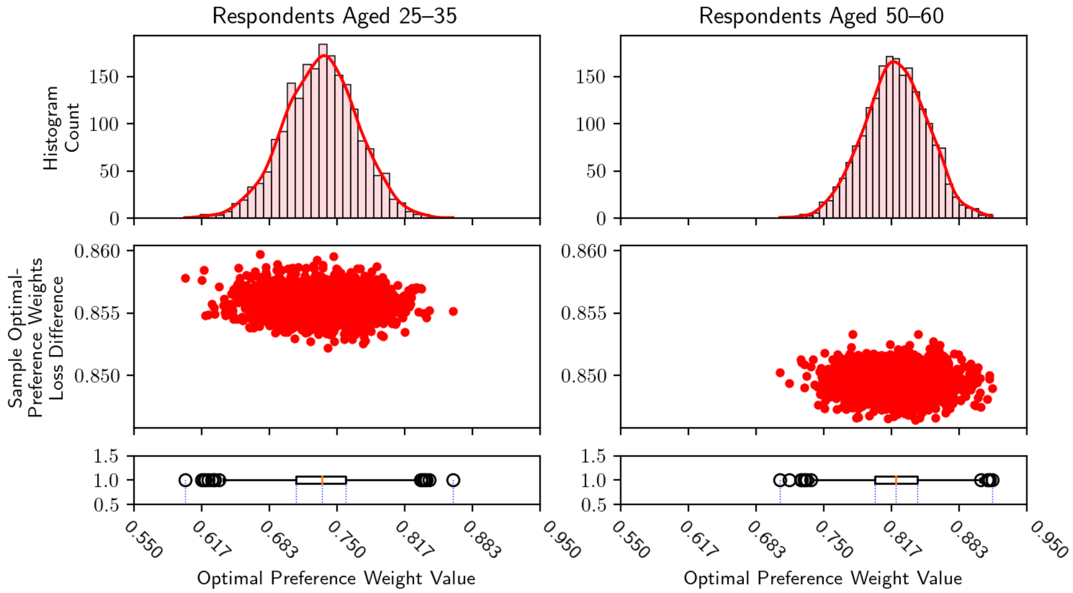
Dimension 40 Preference-Weight Loss-Curve Sampling
"fire-related/heat-related"



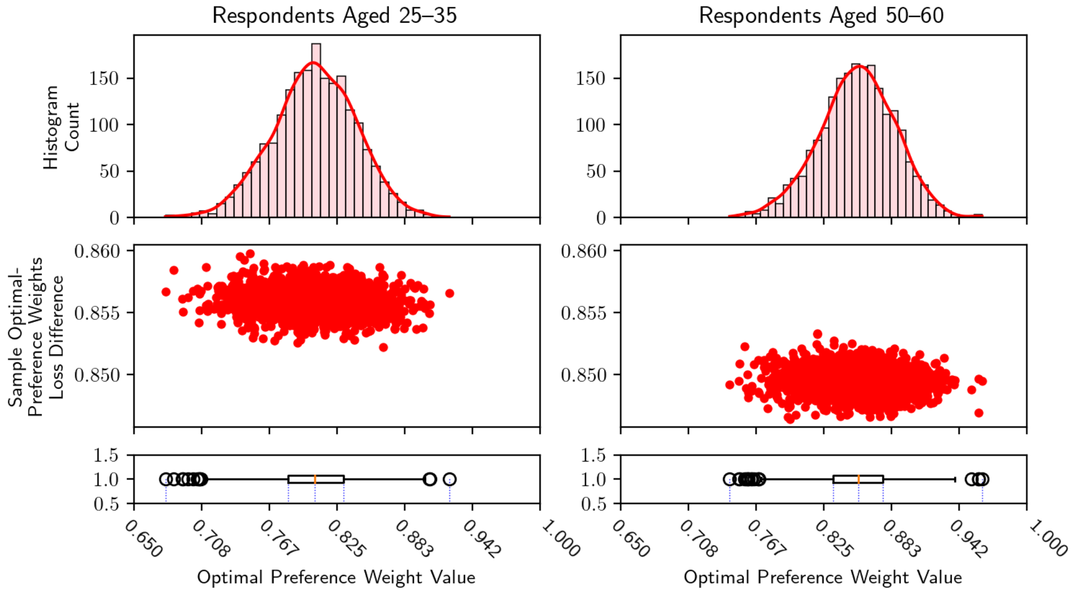
Dimension 41 Preference-Weight Loss-Curve Sampling
"head-related/face-related"



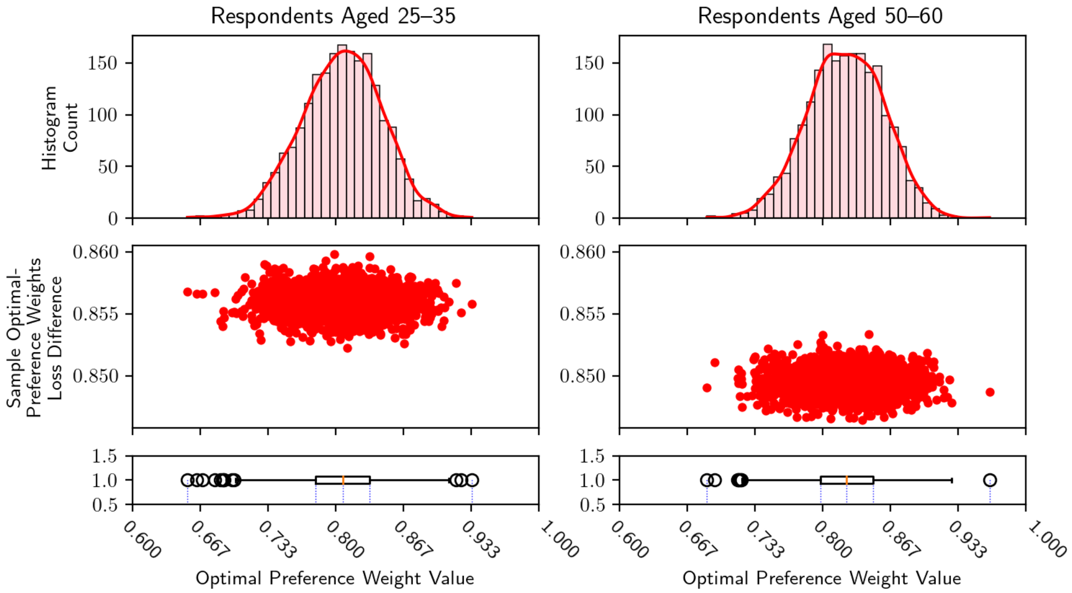
Dimension 42 Preference-Weight Loss-Curve Sampling "beams-related"



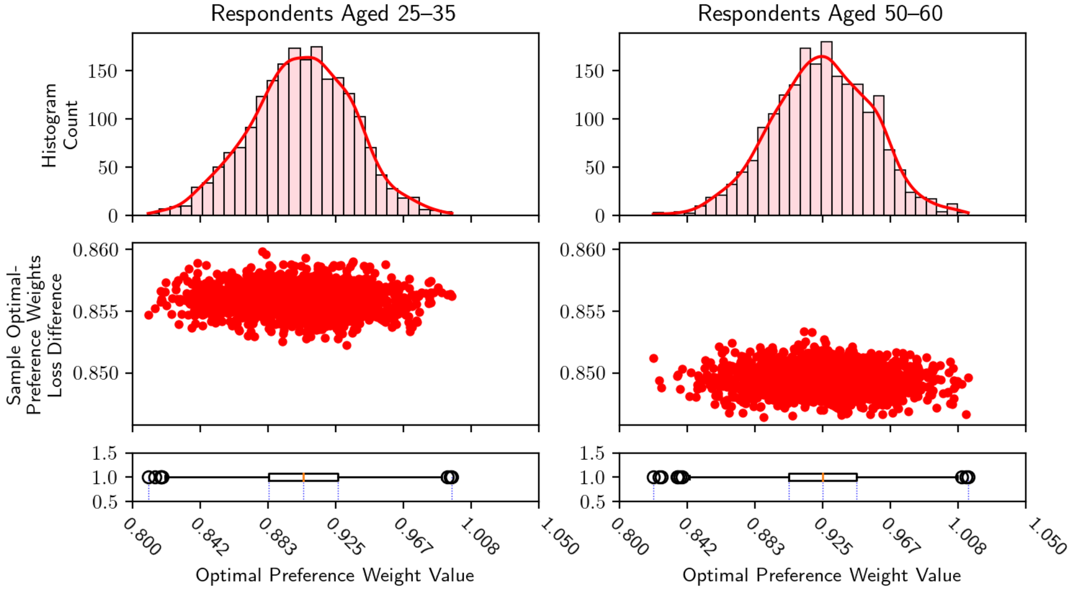
Dimension 43 Preference-Weight Loss-Curve Sampling "seating-related/put things on top"



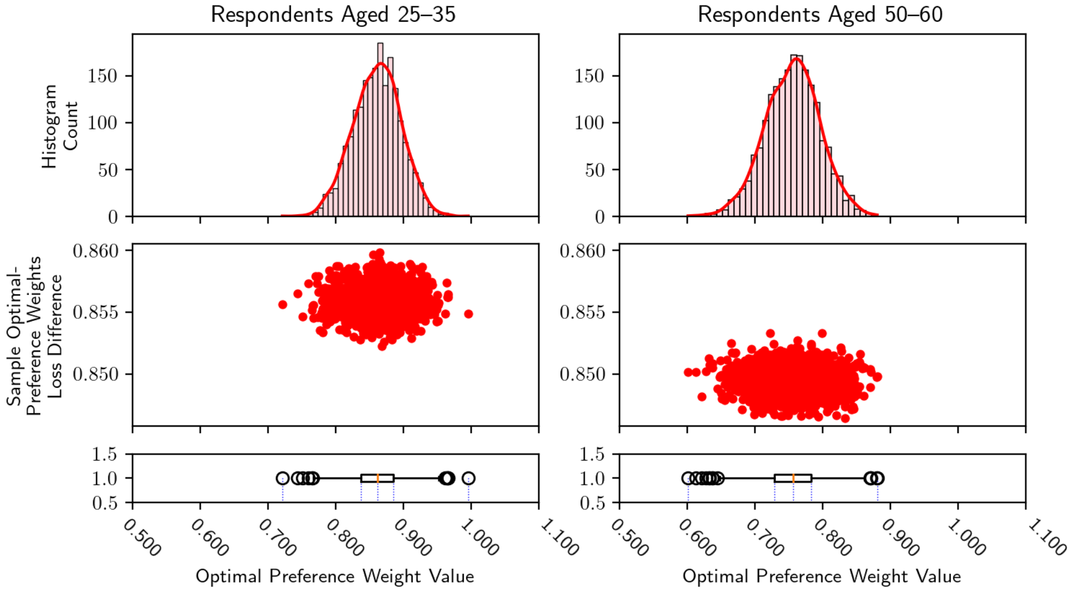
Dimension 44 Preference-Weight Loss-Curve Sampling
"container-related/hollow"



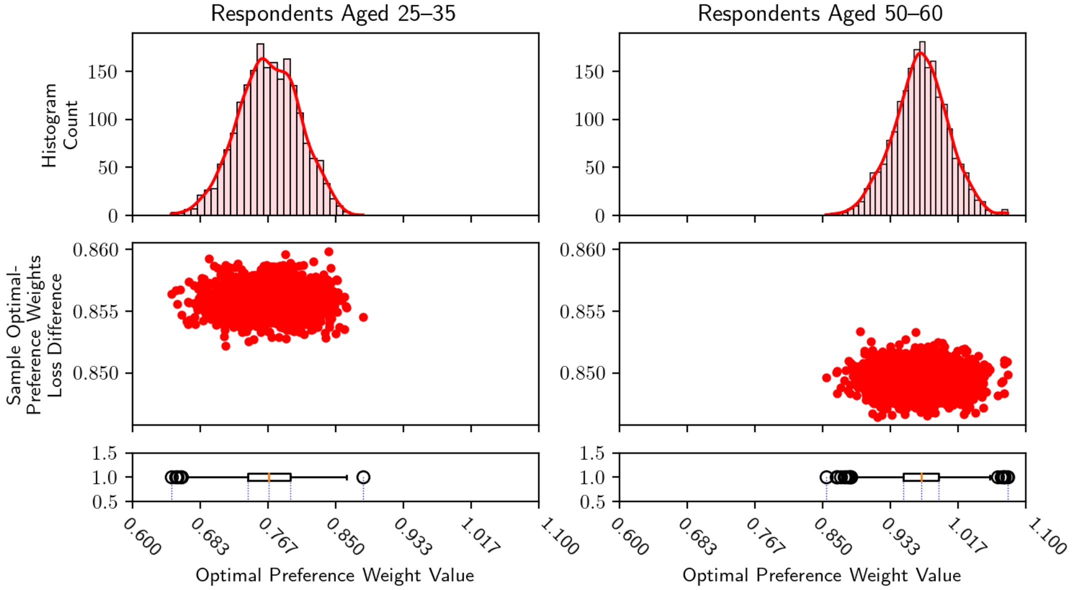
Dimension 45 Preference-Weight Loss-Curve Sampling
"child-related/toy-related"



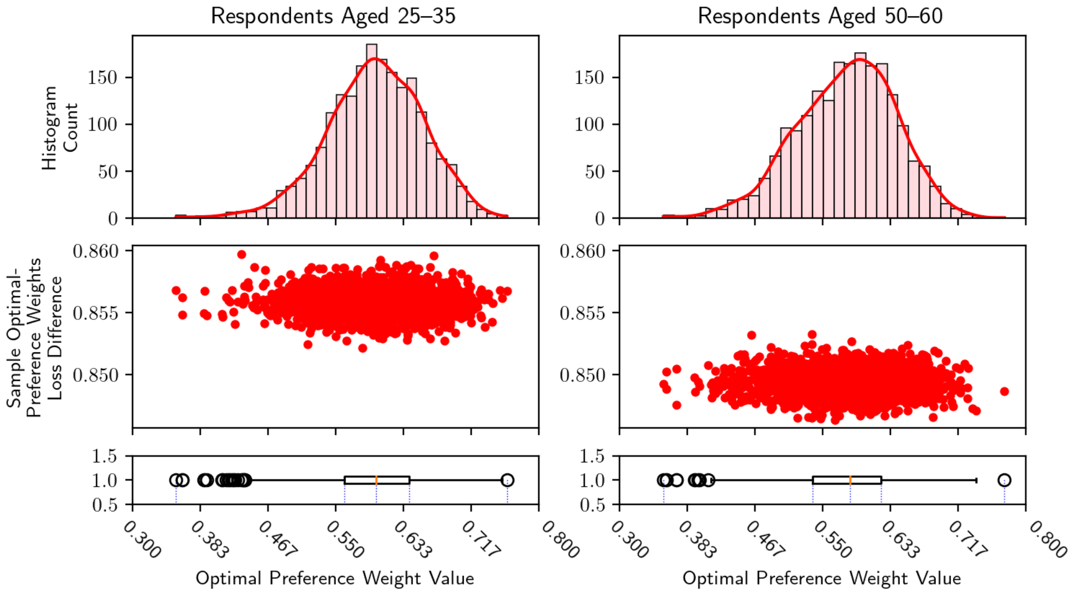
Dimension 46 Preference-Weight Loss-Curve Sampling
"medicine-related"



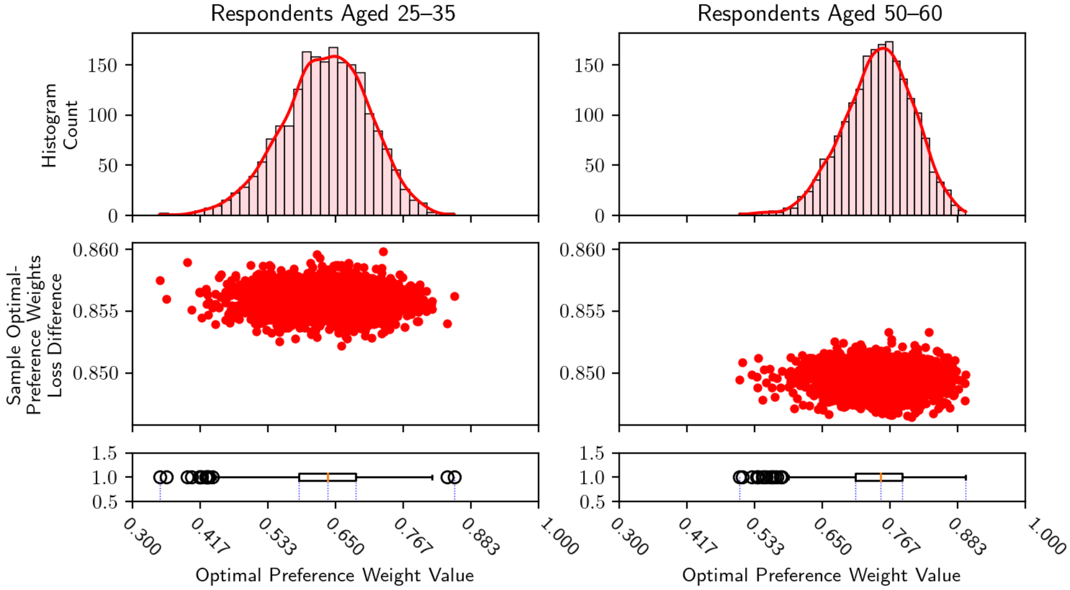
Dimension 47 Preference-Weight Loss-Curve Sampling
"has grating"



Dimension 48 Preference-Weight Loss-Curve Sampling
"handicraft-related"

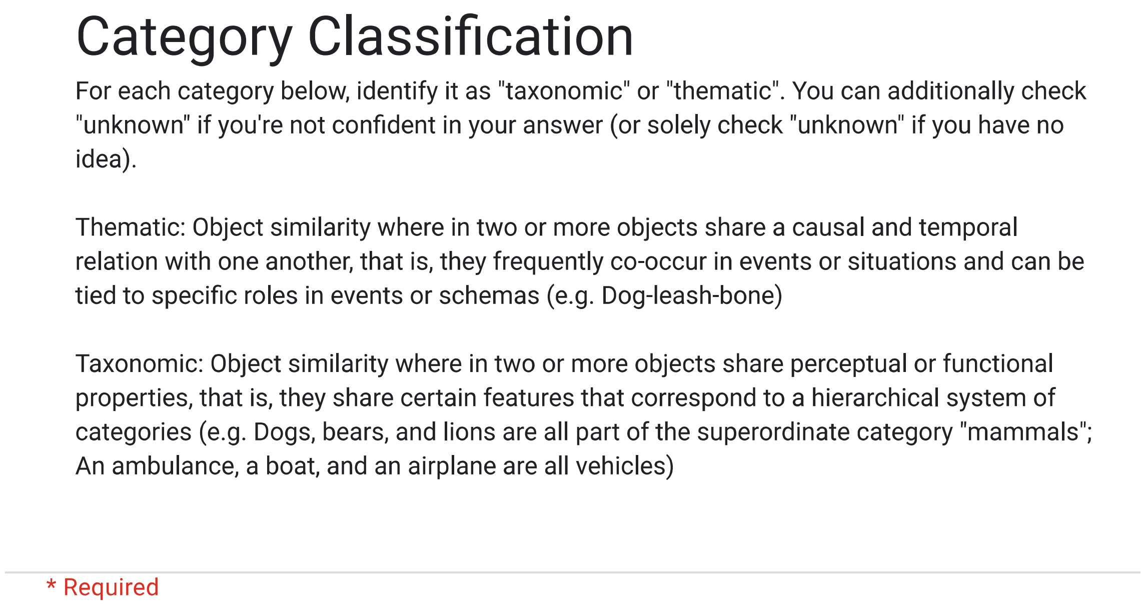


Dimension 49 Preference-Weight Loss-Curve Sampling
"cylindrical/conical"



Appendix B: Taxonomic/Thematic Dimension Labels Survey

We administered the survey for classifying the Hebart et al. model dimensions as being heavily taxonomic or thematic in nature (see Section 4.4.5.2) using Google Forms. A printout of the survey is provided below, although note that all respondents answered with the equivalent online version. The table of results for this form is given in Appendix C.



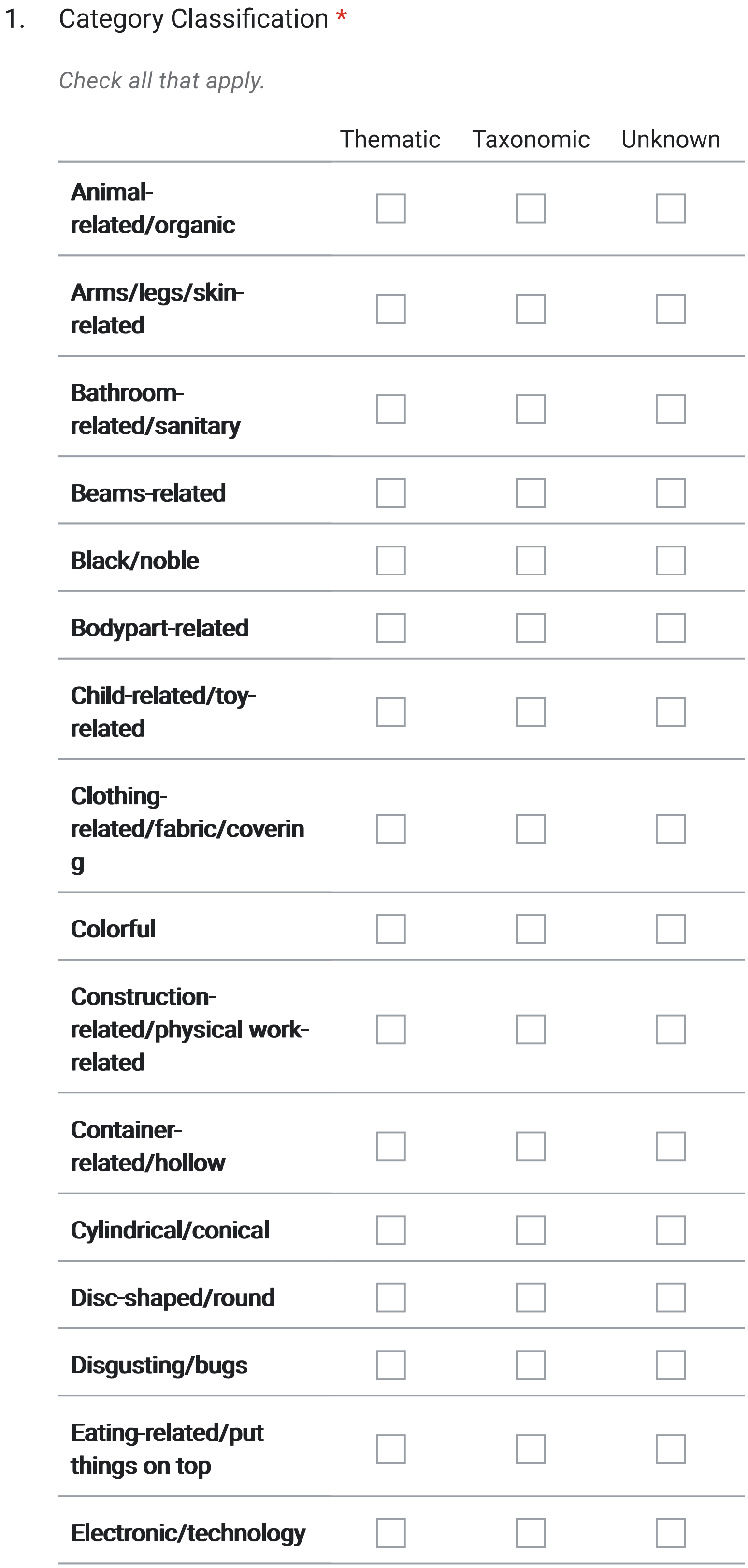
Category Classification

For each category below, identify it as "taxonomic" or "thematic". You can additionally check "unknown" if you're not confident in your answer (or solely check "unknown" if you have no idea).

Thematic: Object similarity where in two or more objects share a causal and temporal relation with one another, that is, they frequently co-occur in events or situations and can be tied to specific roles in events or schemas (e.g. Dog-leash-bone)

Taxonomic: Object similarity where in two or more objects share perceptual or functional properties, that is, they share certain features that correspond to a hierarchical system of categories (e.g. Dogs, bears, and lions are all part of the superordinate category "mammals"; An ambulance, a boat, and an airplane are all vehicles)

* Required

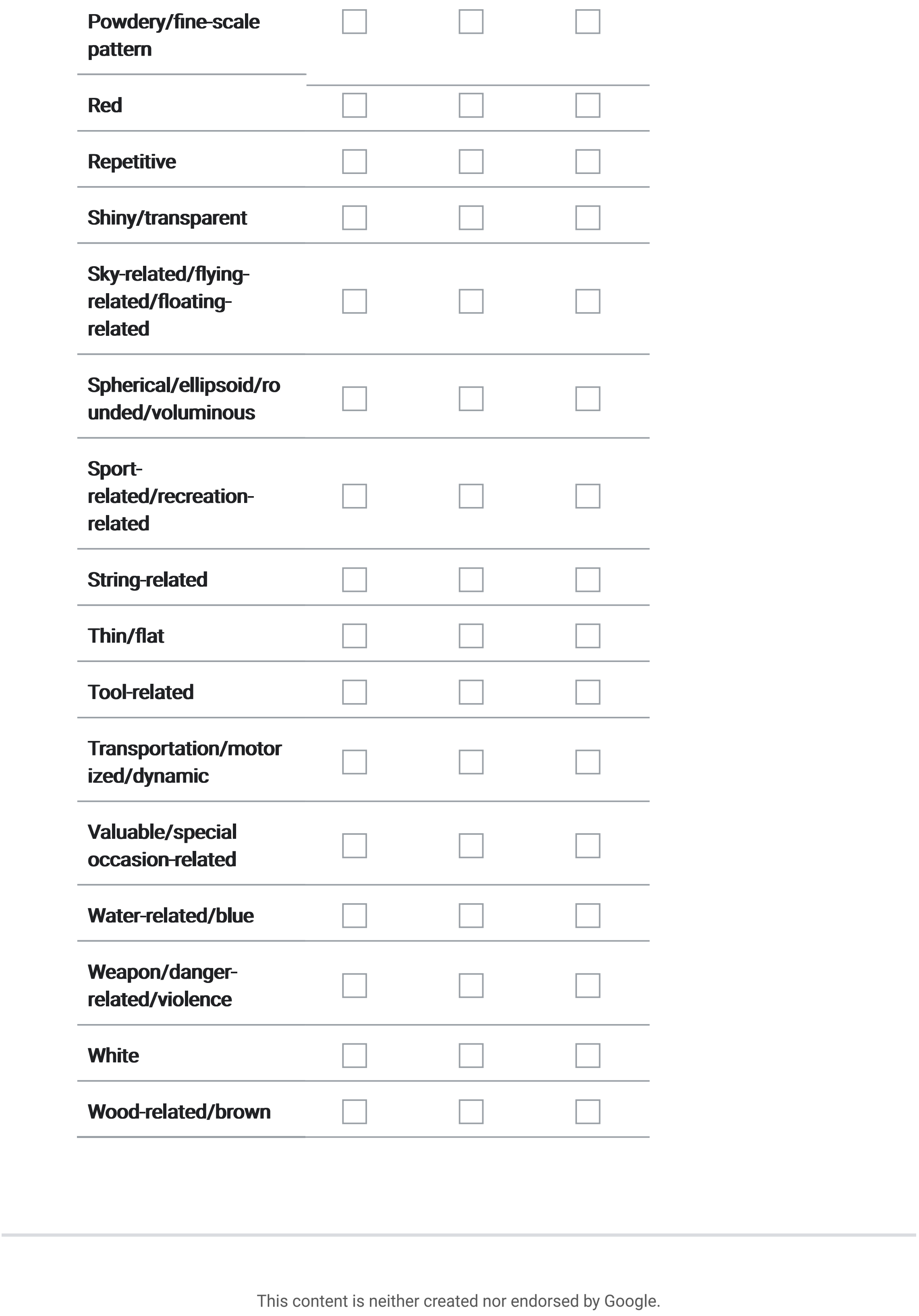


1. Category Classification *

Check all that apply.

	Thematic	Taxonomic	Unknown
Animal-related/organic	□	□	□
Arms/legs/skin-related	□	□	□
Bathroom-related/sanitary	□	□	□
Beams-related	□	□	□
Black/noble	□	□	□
Bodypart-related	□	□	□
Child-related/toy-related	□	□	□
Clothing-related/fabric/covering	□	□	□
Colorful	□	□	□
Construction-related/physical work-related	□	□	□
Container-related/hollow	□	□	□
Cylindrical/conical	□	□	□
Disc-shaped/round	□	□	□
Disgusting/bugs	□	□	□
Eating-related/put things on top	□	□	□
Electronic/technology	□	□	□

Feminine (stereotypically)/decorative	□	□	□
Fire-related/heat-related	□	□	□
Flat/patterned	□	□	□
Fluid-related/drink-related	□	□	□
Food-related/eating-related/kitchen-related	□	□	□
Furniture-related/household-related/artifact	□	□	□
Handicraft-related	□	□	□
Has grating	□	□	□
Head-related/face-related	□	□	□
Long/thin	□	□	□
Made of metal/artificial/hard	□	□	□
Many small things/coarse pattern	□	□	□
Medicine-related	□	□	□
Musical instrument-related/noise-related	□	□	□
Outdoors-related	□	□	□
Paper-related/thin/flat/text-related	□	□	□
Plant-related/green	□	□	□



Powdery/fine-scale pattern	□	□	□
Red	□	□	□
Repetitive	□	□	□
Shiny/transparent	□	□	□
Sky-related/flying-related/floating-related	□	□	□
Spherical/ellipsoid/rounded/voluminous	□	□	□
Sport-related/recreation-related	□	□	□
String-related	□	□	□
Thin/flat	□	□	□
Tool-related	□	□	□
Transportation/motorized/dynamic	□	□	□
Valuable/special occasion-related	□	□	□
Water-related/blue	□	□	□
Weapon/danger-related/violence	□	□	□
White	□	□	□
Wood-related/brown	□	□	□

This content is neither created nor endorsed by Google.

Appendix C: Survey Results and Taxonomic/Thematic Dimension Labels

After administering the survey in Appendix B, we used the results to assign taxonomic, thematic, or unknown labels to each dimension. Consider the margin of responses achieved by subtracting the number of "taxonomic" votes for a dimension from the number of "thematic" votes. Dimensions where this margin was 5 or more (in other words, where there were at least 5 more respondents indicating it was more thematic than taxonomic) were labelled as being thematic dimensions. Correspondingly, dimensions where this value was -5 or more were labelled as being taxonomic. See Section 4.4.5.2 for more details.

Category Classification	Taxo.	Them.	Unkn.	Taxo-Thema Margin	Classification ($\|n\|$>4?)	% Taxo.	% Them.
Animal-related/organic	5	4	1	-1	unknown	0.4545	0.4545
Arms/legs/skin-related	8	1	1	-7	taxonomic	0.7273	0.7273
Bathroom-related/sanitary	0	9	0	9	thematic	0	0
Beams-related	5	0	4	-5	taxonomic	0.4545	0.4545
Black/noble	2	3	4	1	unknown	0.1818	0.1818
Bodypart-related	6	3	1	-3	unknown	0.5455	0.5455
Child-related/toy-related	0	10	0	10	thematic	0	0
Clothing-related/fabric/covering	9	1	0	-8	taxonomic	0.8182	0.8182
Colorful	5	3	2	-2	unknown	0.4545	0.4545
Construction-related/physical work-related	2	8	0	6	thematic	0.1818	0.1818
Container-related/hollow	8	2	0	-6	taxonomic	0.7273	0.7273
Cylindrical/conical	8	2	0	-6	taxonomic	0.7273	0.7273
Disc-shaped/round	8	2	0	-6	taxonomic	0.7273	0.7273
Disgusting/bugs	2	7	1	5	thematic	0.1818	0.1818
Eating-related/put things on top	0	4	5	4	unknown	0	0
Electronic/technology	8	2	0	-6	taxonomic	0.7273	0.7273
Feminine (stereotypically)/decorative	2	7	1	5	thematic	0.1818	0.1818

Table C.1 continued from previous page

Category Classification	Taxo.	Them.	Unkn.	Taxo-Thema Margin	Classification (|n|>4?)	% Taxo.	% Them.
Fire-related/heat-related	7	3	0	-4	unknown	0.6364	0.6364
Flat/patterned	6	2	1	-4	unknown	0.5455	0.5455
Fluid-related/drink-related	6	3	0	-3	unknown	0.5455	0.5455
Food-related/eating-related/kitchen-related	1	9	0	8	thematic	0.0909	0.0909
Furniture-related/household-related/artifact	1	9	0	8	thematic	0.0909	0.0909
Handicraft-related	2	6	2	4	unknown	0.1818	0.1818
Has grating	5	1	4	-4	unknown	0.4545	0.4545
Head-related/face-related	5	5	0	0	unknown	0.4545	0.4545
Long/thin	7	3	0	-4	unknown	0.6364	0.6364
Made of metal/artificial/hard	7	3	0	-4	unknown	0.6364	0.6364
Many small things/coarse pattern	8	1	1	-7	taxonomic	0.7273	0.7273
Medicine-related	3	5	2	2	unknown	0.2727	0.2727
Musical instrument-related/noise-related	3	7	0	4	unknown	0.2727	0.2727
Outdoors-related	1	7	2	6	thematic	0.0909	0.0909
Paper-related/thin/flat/text-related	7	2	0	-5	taxonomic	0.6364	0.6364
Plant-related/green	6	4	0	-2	unknown	0.5455	0.5455

Table C.1 continued from previous page

Category Classification	Taxo.	Them.	Unkn.	Taxo-Thema Margin	Classification ($\|n\|$>4?)	% Taxo.	% Them.
Powdery/fine-scale pattern	6	2	2	-4	unknown	0.5455	0.5455
Red	6	2	2	-4	unknown	0.5455	0.5455
Repetitive	5	2	3	-3	unknown	0.4545	0.4545
Shiny/transparent	7	2	0	-5	taxonomic	0.6364	0.6364
Sky-related/flying-related/floating-related	3	7	0	4	unknown	0.2727	0.2727
Spherical/ellipsoid/rounded/voluminous	9	1	0	-8	taxonomic	0.8182	0.8182
Sport-related/recreation-related	1	9	0	8	thematic	0.0909	0.0909
String-related	7	1	2	-6	taxonomic	0.6364	0.6364
Thin/flat	8	2	0	-6	taxonomic	0.7273	0.7273
Tool-related	5	3	2	-2	unknown	0.4545	0.4545
Transportation/motorized/dynamic	6	4	0	-2	unknown	0.5455	0.5455
Valuable/special occasion-related	1	9	0	8	thematic	0.0909	0.0909
Water-related/blue	3	5	0	2	unknown	0.2727	0.2727
Weapon/danger-related/violence	1	7	0	6	thematic	0.0909	0.0909
White	6	2	2	-4	unknown	0.5455	0.5455
Wood-related/brown	5	5	0	0	unknown	0.4545	0.4545

Appendix D: Adult/Child Taxonomic/Thematic McNemar Results

This appendix has the full results for Experiment 1 of Chapter 4—Section 4.6 only gives significant results. Experiment 1 tested whether children were more likely to have different responses than adults due to taxonomic factors, or due to thematic factors. It more specifically tested this for different levels of strength of these factors. Further discussion of these results can be found in Section 4.6.2.

Confusion Range	p-value	Statistic	Contingency Table Th. Sw. F T Ta. Sw. F T	Taxonomic Swaps	Thematic Swaps	Significant Thematic Priority?	$p < 0.05$	$p < 0.1$
all	0.101	2.694	F 339 77 T 57 22	79	99	FALSE	F	F
0.0–0.1	0.077	3.125	F 39 7 T 1 1	2	8	TRUE	F	T
0.0–0.2	0.034	4.5	F 71 14 T 4 4	8	18	TRUE	T	T
0.0–0.3	0.091	2.857	F 101 23 T 12 7	19	30	TRUE	F	T
0.0–0.4	0.243	1.362	F 130 28 T 19 11	30	39	FALSE	F	F
0.0–0.5	0.435	0.61	F 157 33 T 26 15	41	48	FALSE	F	F
0.0–0.6	0.483	0.493	F 190 40 T 33 15	48	55	FALSE	F	F
0.0–0.7	0.51	0.434	F 223 45 T 38 17	55	62	FALSE	F	F
0.0–0.8	0.223	1.485	F 253 55 T 42 18	60	73	FALSE	F	F
0.0–0.9	0.111	2.535	F 280 66 T 48 21	69	87	FALSE	F	F

Table D.1 continued from previous page

Confusion Range	p-value	Statistic	Contingency Table Th. Sw. F T Ta. Sw. F T	Taxonomic Swaps	Thematic Swaps	Significant Thematic Priority?	$p < 0.05$	$p < 0.1$
0.0–1.0	0.124	2.369	F 298 | 70 T 52 | 22	74	92	FALSE	F	F
0.0–1.1	0.127	2.331	F 312 | 71 T 53 | 22	75	93	FALSE	F	F
0.0–1.2	0.093	2.82	F 329 | 74 T 54 | 22	76	96	TRUE	F	T
0.0–10.0	0.101	2.694	F 339 | 77 T 57 | 22	79	99	FALSE	F	F
0.1–0.2	0.343	0.9	F 32 | 7 T 3 | 3	6	10	FALSE	F	F
0.1–0.3	0.441	0.593	F 62 | 16 T 11 | 6	17	22	FALSE	F	F
0.1–0.4	0.749	0.103	F 91 | 21 T 18 | 10	28	31	FALSE	F	F
0.1–0.5	1	0	F 118 | 26 T 25 | 14	39	40	FALSE	F	F
0.1–0.6	1	0	F 151 | 33 T 32 | 14	46	47	FALSE	F	F

Table D.1 continued from previous page

Confusion Range	p-value	Statistic	Contingency Table Th. Sw. F T Ta. Sw. F T	Taxonomic Swaps	Thematic Swaps	Significant Thematic Priority?	$p < 0.05$	$p < 0.1$
0.1–0.7	1	0	F 184 | 38 T 37 | 16	53	54	FALSE	F	F
0.1–0.8	0.525	0.404	F 214 | 48 T 41 | 17	58	65	FALSE	F	F
0.1–0.9	0.285	1.142	F 241 | 59 T 47 | 20	67	79	FALSE	F	F
0.1–1.0	0.303	1.061	F 259 | 63 T 51 | 21	72	84	FALSE	F	F
0.1–1.1	0.307	1.043	F 273 | 64 T 52 | 21	73	85	FALSE	F	F
0.1–1.2	0.235	1.408	F 290 | 67 T 53 | 21	74	88	FALSE	F	F
0.1–10.0	0.247	1.341	F 300 | 70 T 56 | 21	77	91	FALSE	F	F
0.2–0.3	1	0	F 30 | 9 T 8 | 3	11	12	FALSE	F	F
0.2–0.4	1	0	F 59 | 14 T 15 | 7	22	21	FALSE	F	F

Table D.1 continued from previous page

Confusion Range	p-value	Statistic	Contingency Table Th. Sw. F T Ta. Sw. F T	Taxonomic Swaps	Thematic Swaps	Significant Thematic Priority?	$p < 0.05$	$p < 0.1$
0.2–0.5	0.755	0.098	F 86 19 T 22 11	33	30	FALSE	F	F
0.2–0.6	0.787	0.073	F 119 26 T 29 11	40	37	FALSE	F	F
0.2–0.7	0.804	0.062	F 152 31 T 34 13	47	44	FALSE	F	F
0.2–0.8	0.822	0.051	F 182 41 T 38 14	52	55	FALSE	F	F
0.2–0.9	0.475	0.51	F 209 52 T 44 17	61	69	FALSE	F	F
0.2–1.0	0.492	0.471	F 227 56 T 48 18	66	74	FALSE	F	F
0.2–1.1	0.497	0.462	F 241 57 T 49 18	67	75	FALSE	F	F
0.2–1.2	0.391	0.736	F 258 60 T 50 18	68	78	FALSE	F	F
0.2–10.0	0.403	0.698	F 268 63 T 53 18	71	81	FALSE	F	F

Table D.1 continued from previous page

Confusion Range	p-value	Statistic	Contingency Table Th. Sw. F T Ta. Sw. F T	Taxonomic Swaps	Thematic Swaps	Significant Thematic Priority?	$p < 0.05$	$p < 0.1$
0.3–0.4	0.773	0.083	F 29 | 5 T 7 | 4	11	9	FALSE	F	F
0.3–0.5	0.54	0.375	F 56 | 10 T 14 | 8	22	18	FALSE	F	F
0.3–0.6	0.626	0.237	F 89 | 17 T 21 | 8	29	25	FALSE	F	F
0.3–0.7	0.665	0.188	F 122 | 22 T 26 | 10	36	32	FALSE	F	F
0.3–0.8	0.899	0.016	F 152 | 32 T 30 | 11	41	43	FALSE	F	F
0.3–0.9	0.5	0.456	F 179 | 43 T 36 | 14	50	57	FALSE	F	F
0.3–1.0	0.52	0.414	F 197 | 47 T 40 | 15	55	62	FALSE	F	F
0.3–1.1	0.525	0.404	F 211 | 48 T 41 | 15	56	63	FALSE	F	F
0.3–1.2	0.407	0.688	F 228 | 51 T 42 | 15	57	66	FALSE	F	F

Table D.1 continued from previous page

Confusion Range	p-value	Statistic	Contingency Table Th. Sw. F T Ta. Sw. F T	Taxonomic Swaps	Thematic Swaps	Significant Thematic Priority?	$p < 0.05$	$p < 0.1$
0.3–10.0	0.421	0.646	F 238 54 T 45 15	60	69	FALSE	F	F
0.4–0.5	0.773	0.083	F 27 5 T 7 4	11	9	FALSE	F	F
0.4–0.6	0.845	0.038	F 60 12 T 14 4	18	16	FALSE	F	F
0.4–0.7	0.868	0.028	F 93 17 T 19 6	25	23	FALSE	F	F
0.4–0.8	0.671	0.18	F 123 27 T 23 7	30	34	FALSE	F	F
0.4–0.9	0.328	0.955	F 150 38 T 29 10	39	48	FALSE	F	F
0.4–1.0	0.356	0.853	F 168 42 T 33 11	44	53	FALSE	F	F
0.4–1.1	0.362	0.831	F 182 43 T 34 11	45	54	FALSE	F	F
0.4–1.2	0.267	1.235	F 199 46 T 35 11	46	57	FALSE	F	F

Table D.1 continued from previous page

Confusion Range	p-value	Statistic	Contingency Table Th. Sw. F T Ta. Sw. F T	Taxonomic Swaps	Thematic Swaps	Significant Thematic Priority?	$p < 0.05$	$p < 0.1$
0.4–10.0	0.284	1.149	F 209 49 T 38 11	49	60	FALSE	F	F
0.5–0.6	0.789	0.071	F 33 7 T 7 0	7	7	FALSE	F	F
0.5–0.7	0.838	0.042	F 66 12 T 12 2	14	14	FALSE	F	F
0.5–0.8	0.417	0.658	F 96 22 T 16 3	19	25	FALSE	F	F
0.5–0.9	0.178	1.818	F 123 33 T 22 6	28	39	FALSE	F	F
0.5–1.0	0.208	1.587	F 141 37 T 26 7	33	44	FALSE	F	F
0.5–1.1	0.215	1.538	F 155 38 T 27 7	34	45	FALSE	F	F
0.5–1.2	0.149	2.087	F 172 41 T 28 7	35	48	FALSE	F	F
0.5–10.0	0.166	1.92	F 182 44 T 31 7	38	51	FALSE	F	F

Table D.1 continued from previous page

Confusion Range	p-value	Statistic	Contingency Table			Taxonomic Swaps	Thematic Swaps	Significant Thematic Priority?	$p < 0.05$	$p < 0.1$
			Th. Sw.	F	T					
			Ta. Sw.							
0.6–0.7	0.752	0.1	F	33	5	7	7	FALSE	F	F
			T	5	2					
0.6–0.8	0.307	1.042	F	63	15	12	18	FALSE	F	F
			T	9	3					
0.6–0.9	0.118	2.439	F	90	26	21	32	FALSE	F	F
			T	15	6					
0.6–1.0	0.153	2.041	F	108	30	26	37	FALSE	F	F
			T	19	7					
0.6–1.1	0.161	1.961	F	122	31	27	38	FALSE	F	F
			T	20	7					
0.6–1.2	0.106	2.618	F	139	34	28	41	FALSE	F	F
			T	21	7					
0.6–10.0	0.124	2.361	F	149	37	31	44	FALSE	F	F
			T	24	7					
0.7–0.8	0.181	1.786	F	30	10	5	11	FALSE	F	F
			T	4	1					
0.7–0.9	0.072	3.226	F	57	21	14	25	TRUE	F	T
			T	10	4					

Table D.1 continued from previous page

Confusion Range	p-value	Statistic	Contingency Table Th. Sw. F T Ta. Sw. F T	Taxonomic Swaps	Thematic Swaps	Significant Thematic Priority?	$p < 0.05$	$p < 0.1$
0.7–1.0	0.109	2.564	F 75 25 T 14 5	19	30	FALSE	F	F
0.7–1.1	0.118	2.439	F 89 26 T 15 5	20	31	FALSE	F	F
0.7–1.2	0.074	3.2	F 106 29 T 16 5	21	34	TRUE	F	T
0.7–10.0	0.093	2.824	F 116 32 T 19 5	24	37	TRUE	F	T
0.8–0.9	0.332	0.941	F 27 11 T 6 3	9	14	FALSE	F	F
0.8–1.0	0.424	0.64	F 45 15 T 10 4	14	19	FALSE	F	F
0.8–1.1	0.441	0.593	F 59 16 T 11 4	15	20	FALSE	F	F
0.8–1.2	0.281	1.161	F 76 19 T 12 4	16	23	FALSE	F	F
0.8–10.0	0.324	0.973	F 86 22 T 15 4	19	26	FALSE	F	F

Table D.1 continued from previous page

Confusion Range	p-value	Statistic	Contingency Table Th. Sw. F T Ta. Sw. F T	Taxonomic Swaps	Thematic Swaps	Significant Thematic Priority?	$p < 0.05$	$p < 0.1$
0.9–1.0	0.724	0.125	F 18 4 T 4 1	5	5	FALSE	F	F
0.9–1.1	0.752	0.1	F 32 5 T 5 1	6	6	FALSE	F	F
0.9–1.2	0.789	0.071	F 49 8 T 6 1	7	9	FALSE	F	F
0.9–10.0	0.823	0.05	F 59 11 T 9 1	10	12	FALSE	F	F
1.0–1.1	0.48	0.5	F 14 1 T 1 0	1	1	FALSE	F	F
1.0–1.2	0.683	0.167	F 31 4 T 2 0	2	4	FALSE	F	F
1.0–10.0	0.773	0.083	F 41 7 T 5 0	5	7	FALSE	F	F
1.1–1.2	0.617	0.25	F 17 3 T 1 0	1	3	FALSE	F	F
1.1–10.0	0.752	0.1	F 27 6 T 4 0	4	6	FALSE	F	F

Table D.1 continued from previous page

Confusion Range	p-value	Statistic	Contingency Table Th. Sw. F T Ta. Sw. F T	Taxonomic Swaps	Thematic Swaps	Significant Thematic Priority?	$p < 0.05$	$p < 0.1$
1.2–10.0	0.683	0.167	F 10 3 T 3 0	3	3	FALSE	F	F

Appendix E: Code Repositories

Code for both projects of this thesis can be found on GitHub.

The code for Chapter 3, which examines adult age and object-similarity preferences, is located at https://github.com/fyshelab/adult-likeness-preference-weights-project.

The code for Chapter 4, which focuses on six-year-olds' taxonomic and thematic reasoning, is located at https://github.com/fyshelab/child-adult-taxonomic-thematic-project.