THE UNIVERSITY OF CHICAGO

IMPACT OF MENTAL REPRESENTATION ON CONSUMER BEHAVIORS: IMPLICATIONS FOR MENTAL BUDGETING AND PREDICTION ALGORITHM PREFERENCES

A DISSERTATION SUBMITTED TO

THE FACULTY OF THE UNIVERSITY OF CHICAGO

BOOTH SCHOOL OF BUSINESS

IN CANDIDACY FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

LIN FEI

CHICAGO, ILLINOIS

JUNE 2023

Table of Contents

List of Tables

List of Figures

Acknowledgements

I am extremely grateful to my co-advisors Dan Bartels and Berkeley Dietvorst, for showing me how rigorous and scientific research is and should be, for challenging me to articulate and explain my thinking, and for always giving thoughtful advice and words of encouragement. I am also incredibly thankful to Luxi Shen, whose mentorship and friendship have been essential to my research experience and personal development, as well as Reid Hastie, who has always been so generous with his time and wisdom and has always given out careful thoughts to my projects. I deeply appreciate my fellow cohort Danny Katz and Minkwang Jang who are always there for their peer, as well as the HOCAG group which is always an encouraging and fun environment to discuss research.

Finally, I am extremely grateful for Ignatius Liu, who has been incredibly supportive of me as a researcher and a person. Your help and your company have contributed much to this dissertation and this wonderful journey.

Overview

People constantly encode and represent the information around them, and they make various inferences, judgments, and decisions based on their representation. The dissertation aims to connect mental representations to consumer behaviors. Investigating how consumers represent their choices and contexts allows us to generate new predictions about their behaviors and provide novel insights into the cognitive processes under these enriched contexts. In Chapter 1, I explore the relationship between the representation of concepts and spending behaviors. In Chapter 2, I explore how people represent prediction events and consequently how it influences their prediction algorithm choices. Additional norming studies and detailed study stimuli are available in Supplemental Files: Web Appendix.

Introduction

People constantly encode and represent the information around them, and they make various inferences, judgement, and decisions based on their representation. For example, people could use groups of exemplars or prototypes to represent concepts, could leverage maps or networks to represent relationships, and could adopt mental models or scripts to represent processes. The study of mental representations is foundational in understanding the human mind because it is central to how humans process and make use of information.

Understanding how people represent the world around them has much merit particularly in consumer contexts (Bartels and Johnson 2015). A coherent theory on people's representation can help highlight the potential common ground of seemingly distinct findings and unify multiple established phenomena. For example, how people represent monetary values can provide an explanation to concave utility function, loss aversion, and hyperbolic discounting (Stewart, Charter, and Brown 2006). In addition, understanding consumer's representations allows us to further predict their reactions and behaviors in contexts like new product launches (Moreau, Markman, and Lehmann 2001), product bundles, and service failures. In applying representation to the study of consumer behaviors, we simultaneously learn about the behavior and the thinking behind it.

More generally, identifying the role of representation and cognition complements the common economic approach to studying consumer behavior. The study of consumer behaviors has long been focused on how people's behaviors and judgments deviate from those that are economically optimal (Thaler 2015). The cognitive approach on the other hand puts much emphasis on why people are behaving the way they are and where the phenomena come from. Combining the two approaches provides a better understanding of why consumers make

decisions in particular ways and generates predictions on when consumers will deviate from the economic prediction.

Studying representation with these enriched environments and complex decision problems in consumer contexts also provides new insights into the human mind. Consumers face some of the most complicated decisions such as purchasing a house, and the study of these choices invites new theories on problem solving and decision making (Payne, Bettman, and Johnson 1993). For example, the study of brand and product categorization expands our understanding on how a category prompts the objects in it (Nedungadi 1990), and the study of context effects in product choices (Huber, Payne, and Puto 1982) has inspired much theorization on how contexts shape decisions (Trueblood et al. 2014).

Finally, connecting representation with behavior allows us to identify contexts where behaviors pertaining to daily lives, spending, and decisions, and thinking are related. In this dissertation, I aim to identify two of these contexts. Specifically, I found that representations influence people's behaviors in how they manage their budgets and how they choose prediction algorithms, in ways that are distinct from predictions of simple normative models.

In Chapter 1, I explore the relationship between the representation of concepts and spending behaviors. Through 4 sets of studies, I find that when consumers deviate from budget on one item, they are more likely to adjust spending for items closer in representation than further. In Study 1, I recover the taxonomy that people use to represent common expenditures with a successive pile-sort paradigm. Study 2 found that recovered taxonomy predicts people's self-reported spending adjustment, while Study 3 and 4 found that the taxonomy predicts people's consequential choices and actual grocery purchase behavior. This chapter provides evidence that representation influences spendings, even when money is theoretically fungible.

In Chapter 2, I explore how people represent prediction events and consequently how they choose prediction algorithms. Through 4 sets of studies, I found that people prefer to use prediction algorithms that replicate the same process of the even being predicted, which implies that the mental models of a prediction event include the exact process through which the outcome is generated. I found that when predicting both simple but random tasks (Study 1) and real-world outcomes (Study 2), people prefer algorithms that replicate the process of the prediction event, even when the algorithms might prefer worse. Study 3 provides a boundary of this preference, while Study 4 designs and tests an intervention based on this preference. This chapter highlights how people's representation changes consumers' preferences and choices.

Chapter 1: Consumers' Mental Representation of Expenditures: Implications for Spending and Savings Decisions

Abstract

People's mental representation of expenditures is crucial to how they budget. We propose that much like how people represent natural objects, people represent expenditures in a hierarchical taxonomy. Across seven studies, we found evidence of a hierarchical representation of expenditures. We first recover people's mental representation using a successive pile-sort method that asks people to form hierarchies of categories with common expenditures (e.g., rent, dining out, etc.). We found that there is consensus in people's hierarchical representations of expenditures and that their representations are relatively stable over time. Further, we found that people's adjustment in their spending behavior can be predicted by the distance between items in their representation. Specifically, when people overspent on an item, they were more likely to spontaneously adjust spending for items closer in representation than further. We examine this spontaneous adjustment behavior using both lab studies and field data with over 7-million grocery shopping trips. The findings highlight the connection between mental representation and consumer behavior, and emphasize the importance of studying cognitive representation in the context of consumption.

Consumers often maintain budgets. That is, they set restrictions on how much they wish to spend, and they behave as though money is nonfungible (Thaler 1985; 1999) and earmarked for specific functions (e.g., gas, food; Antonides et al. 2011; Hasting and Shapiro 2013, 2018; Heath and Soll 1996; Cheema and Soman 2006; Soman and Cheema 2011; Thaler 1999). For example, consumers might set a budget specifically intended for gas. When gas prices drops, consumers, who now have a gas-budget surplus, may use it to purchase premium gas (Hastings and Shapiro 2013). In a recent survey, 80% of people reported that they have formally or informally budgeted at some point in their lives, and the practice of budgeting is positively correlated with financial wellbeing (Zhang et al. 2020). Therefore, understanding how budgeting works is essential for understanding consumer financial decision making and consumer welfare.

Research on budgeting mostly theorizes that consumers group expenditures into a single level of budgeting categories (Heath and Soll 1996; Cheema and Soman 2006, Zhang et al. 2020, cf. Henderson and Peterson 1992). Some consumers might group items into relatively general categories such as “Food” and “Entertainment”, while others group to more detailed levels such as “Groceries”, “Dining Out”, and “Movies” (Zhang et al. 2020). In this paper, we propose and test a more complete account that people mentally represent expenditures in hierarchical taxonomies. In other words, people represent expenditures in multiple, nested levels of categories where lower levels of the categories are more specific and higher levels are more general. For example, people might think of cereal as “breakfast food” before they think of it more generally as “food” (Markman, Brendl, and Kim 2007), and hence the “breakfast food” category is nested within the higher-level “food” category (Figure 1).

Consequently, we hypothesize that the taxonomic distance between items is important for people’s spending decisions in situations where they deviate from their budget. We define

taxonomic distance as the level at which the expenditures are categorized together in consumers' taxonomies, and we predict that when people's spending on an item deviates from their budget, they will adjust more on the items that are taxonomically closer than further. This prediction is important to marketers who are interested in understanding the type of products consumers adjust and the degree of adjustment when consumers encounter spending deviations such as holiday splurges. Further, past research that hypothesizes single level budgeting categories cannot account for this adjustment. In this paper, we adopt a method from cognitive anthropology to approximate the represented taxonomy of expenditures. Then, we use the taxonomy to predict how people adjust their spending in various contexts, including real shopping behaviors.

Theoretical Background

Connecting Mental Representation to Mental Accounting

Mental accounting is grounded in the way that people categorize and represent concepts (Henderson and Peterson 1992). Mental accounting refers to the cognitive operations by which prospects are evaluated with respect to some specific "topical" account instead of a "comprehensive" account (Linville and Fischer 1991; Thaler and Johnson 1990; Kahneman and Tversky 1984; Soman 2004; Thaler 1999; Zhang et al. 2020). For example, people are more likely to travel to save $5 on a calculator that originally costs $15 than when it costs $125. This is likely because they evaluate the $5 with respect to the total "topical" expenditure ($15 or $125) but not their total wealth. In doing so, people are associating money with different expenditure concepts, and how people represent the concepts will consequently influence how they spend the resources.

Researchers have connected mental representation with mental accounting (Arkes et al. 1994; Levav and McGraw 2009; Evers, Imas, and Kang 2022) and have provided insights on how people form and use mental accounts. For example, representation of the source of money influences the types of mental accounts that are constructed (Levav and McGraw 2009, Reinholtz, Bartels, and Parker 2015): money associated with negative emotions is less likely to be spent on hedonic purchases, and money associated with brand-specific gift cards is more likely to be spent on brand-typical items. Representation of expenditures also influences the membership of expenditures with respect to a mental account. Expenditures that have a high degree of feature-based (i.e., attribute) similarity are often integrated together as a larger loss (Evers, Imas, and Kang 2022).

Similarly, mental representation has provided much insight into how people mentally budget (Henderson and Peterson 1992; Heath and Soll 1996). If people mentally represent expenditures, for example, they would spontaneously categorize expenditures and organize them in categories when prompted to think about budgets, which is what researchers have found (Cheema and Soman 2006; Sussman and Alter 2012). The expenditures in a category are often similar to each other in some respects: they may share similar features (e.g., electronics) or fulfill similar goals (e.g., entertainment; Felcher, Malaviya, and McGill 2001). Expenditures that are grouped into a budgeting category have graded membership—the typicality of an expenditure within a category can differ (Reinholtz et al. 2015), which is another prediction from research on mental representation. For example, shirts can be more typical of the clothing category than gloves (Heath and Soll 1996).

However, previous research has not connected the hierarchical nature of representation with mental accounting. To the authors' knowledge, only one paper has provided preliminary

tests that people represent mental accounts hierarchically (Henderson and Peterson 1992). Specifically, Henderson and Peterson (1992) mentioned that people have a larger budget for higher-level categories but did not provide predictions on how people make spending decisions when there are hierarchical budgets. However, the researchers did not make concrete predictions on how the hierarchy influences budgeting behavior. In the sections below, we detail the reasoning for proposing a hierarchical representation of expenditures, and we highlight additional contributions of introducing a hierarchical representation in budgeting.

Hierarchical Representation of Expenditures

There are several reasons why we think consumers group expenditures hierarchically. First, research suggests that people represent concepts of the natural world and of their everyday lives in taxonomies (Berlin 1992). People may group cats and dogs into the category "mammals" and then group mammals together with birds and fish into the superordinate category "animal". Organizing concepts at multiple, nested levels is a cognitively efficient way for people to store and access information related to different objects (Collins and Quillian 1969; Smith 1978) and generalize the information to more specific concepts (e.g., a dog inherits the properties of a mammal; Osherson et al. 1990). Research suggests that people hierarchically represent animals (e.g., Lopez et al. 1997), trees (Collins and Quillian 1969; Medin et al. 1997), furniture (Rosch et al. 1976), and clothing (Rosch and Lloyd. 1978), and we suspect that people represent concepts of expenditures in terms of multiple, nested categories as well.

Further, people also represent objects in a way that reflects their experience and their knowledge of the world (Markman 1999; Medin and Atran 2004). For example, the taxonomy for trees that landscapers use tends to reflect differences between trees in shape and appearance, while botanists focus more on the scientific taxonomy. Each group's taxonomy reflects the

knowledge of trees that is most useful for the way that each group typically interacts with and thinks about trees. Similarly, people's interactions with consumer goods likely also shape the categories they use to represent them. For example, grocery items are often subdivided into aisles that serve different functions (breakfast food, frozen meals, paper products, etc.). Further, aisles typically have different sections dedicated to smaller groupings of products (e.g., baking goods aisle may contain both kitchen appliances and flour). As consumers navigate through these shopping environments, they will likely internalize this organizational scheme into their representation.

Finally, previous research on consumer behavior has hinted that people represent expenditures hierarchically. Several studies have suggested that consumers represent snack foods and fast-food brands in taxonomies (Ratneshwar and Shocker 1991; Nedungadi, 1990). Specifically, Ratneshwar and Shocker (1991) recovered a hierarchical structure from their similarity rating data. Nedungadi (1990) investigated the relationship between brand hierarchy and brand recall. In this paper, we extend the view that people represent objects hierarchically and apply it to spending and budgeting.

Some Contributions of a Taxonomy-Based Theory of Mental Budgeting

The notion that expenditures are represented in terms of a taxonomy addresses some limitations in the existing literature on mental budgeting. First, the current budgeting literature often assumes that people have predefined budgeting categories. Papers that investigate the type of budgets people construct often elicit those categories in a top-down fashion, asking directly for the budgeting categories people have in mind (Heath and Soll 1996, Zhang et al. 2020). However, people might come up with the most common categories they budget for (e.g., "food"), and hence the listed budgets might often miss categories like exceptional expenditures

(Sussman and Alter 2012). Relatedly, previous papers on mental budgeting also often prompted people with pre-defined categorical labels when investigating budgeting behaviors (Heath and Soll 1996; Cheema and Soman 2006). For example, participants would be examining their budgets for "food" and for "entertainment", which are researcher-defined categories, when they are deciding how to categorize their expenditure of "dining out". Yet not every constructs these specific budgeting categories (Zhang et al. 2020), and we have limited insights into how people budget in absence of these given categories.

In addition, much of the budgeting literature suggests that consumers have defined expenditure membership with respect to the predefined budgeting categories. Consumers tend to adjust more on items that are typical to these categories (Heath and Soll 1996, Reinholtz et al. 2015), and are worse at evaluating expenditures that are exceptional (Sussman and Alter 2012). However, it is still unclear how consumers define the typicality or exceptionality of an expenditure. Further, consumers often are flexible in the way they categorize expenses (Cheema and Soman 2006) and might categorize a specific expenditure differently depending on the context ("dining out" can belong to "food" or "entertainment"). This suggests that consumers might not have clear categorical membership for expenditures.

A hierarchical representation, on the other hand, puts less emphasis on how expenditures relate to categories and focuses more on how expenditures relate to each other. In this paper, we recover the representation through a bottom-up approach that asks people to group expenditures into a hierarchy. We therefore only argue that the distance in the hierarchy matters for people's budget and do not need to make assumptions on the categories consumers construct or the membership of expenditures. Further, distances in a hierarchical representation can also reveal expenditures that are clustered together. This complements the current literature by providing

insight into the expenditures that are often thought of together and the categories that people naturally construct. With a hierarchical representation, the recovered structure will also reveal the items that are exceptional (e.g., items that are higher up in the hierarchy) or items that can be flexibly categorized (e.g., items that are in the middle of different concepts), which provides additional insight into people's budgeting behaviors.

The ideas in the current manuscript add new, refined predictions about how people adjust their spending when they have overspent or underspent on an item. We first extend the framework of budgeting beyond binary categorical restrictions. Heath and Soll (1996) theorized that consumers set budget categories, and then track and post spending with respect to the category. In this case, spending decisions are binary: when consumers overspent with respect to their budget for a category, they were more likely to adjust their consequent spending on other items in the category than those outside of the category. However, people's purchase behaviors could be influenced by their out-of-category purchases (Chintagunta and Haldar 1998; Manchanda et al. 1999). For example, the purchases of soup and yogurt, which are presumably of different product categories, influence each other (Chintagunta and Halder 1998). Therefore, the framework of budgeting should extend beyond binary categorical restrictions. Our proposed hierarchical representation allows us to examine expenditures at different taxonomic distances beyond binary categorical restrictions. We make a concrete prediction that when a consumer deviates from their budget on a spending, they will adjust more on items that are closer in taxonomic distance than further.

More generally, our paper aims to expand on the theorization of mental accounting beyond the binary distinction of "topical accounts" and "comprehensive accounts" (Kahneman and Tversky 1984; Thaler 1999). Mental accounting is often seen as the behavior that "relates the

consequences of possible choices to a reference level that is determined by the context within which the decision arises" (Thaler 1999). For example, people evaluate a discount of a calculator against the topical account of a calculator purchase rather than against their comprehensive wealth. In contrast, a hierarchical structure implies that there could be multiple nested reference accounts, and people might recruit different reference levels depending on the expenditures they are evaluating. In other words, consumers might be referencing nested mental accounts on a more continuous spectrum than just "topical" and "comprehensive" accounts.

Our prediction not only extends the budgeting literature but also provides insight into consumer research related to cross-product elasticities. Specifically, we predict that after spending on some items, people might adjust down on taxonomically close items due to their budget restraint, resulting in a pattern that is similar to positive cross elasticity. Much examination on cross-product elasticities either recovers the relationship from an observed dataset (Tian, Lautz, Wallis and Lambiotte 2021), or theorizes that the cross-product elasticities were results of substitutes and complements (Koszegi and Matejka 2020). Our theorization suggests an alternative mechanism that cross-product elasticities could emerge from the influence of taxonomic distance on budgeting behaviors. The construct of taxonomic distance is relevant but distinct from substitutability, because it requires people to evaluate their spending with respect to a budget. Further, a close association between items (i.e., a smaller taxonomic distance) might form because of multiple factors. Items closely associated might overlap in their usage contexts (Ratneshwar and Shocker 1991) or related to the same spending goals (Barsalou 1983, 1985). Therefore, people might group items together because the items are substitutes

(e.g., burger and pizza), complements (e.g., shirt and jeans), both (e.g., detergents and soap) or neither (e.g., sunglasses and watch)[1].

We note that the theory of a hierarchical representation in budgeting is an expansion but not an alternative account to the current budgeting literature. We replicate and extend the finding that consumers construct budgeting categories and within a predefined category, expenditure characteristics like typicality matter for spending behaviors. Our paper tests whether consumers also naturally construct categories in a nested fashion, and such hierarchical representation allows us to examine and address open questions within the budgeting literature.

Implications of a Hierarchical Taxonomy

Although we cannot directly examine how people represent expenditures in their minds, we test several implications from our proposal that people represent expenditures hierarchically. First, we can test if consumers have a shared understanding that some items are more closely related than others. For example, we could evaluate whether consumers share the understanding that shampoo and sunscreen are grouped together at a more specific, lower level than shampoo and movie tickets. We test this commonality between people's understanding by testing whether there is statistical consensus among their representations. The term "consensus" is adopted from the Cultural Consensus Model (CCM; Romney, Weller, and Batchelder 1986, Medin et al. 1997) to suggest statistical agreement or commonality across individuals' representations. The CCM establishes consensus by testing whether participants' representations are positively correlated, and whether the first latent root of the interpersonal correlation matrix is large (Weller, 2007). A consensus between participants would suggest not only that people can represent expenditures in

[1] These examples are from our Study 2c, where we collected substitutability and complementarity ratings for all the product pairs that share the smallest taxonomic distance. The details of the collection method are reported in Study 2c.

a taxonomy, but also that this representation is relatively natural to majority of people. Past research has found that members of a culture have a consensual understanding of animal categories (Lopez et al. 1997) and consumers largely agree on product hierarchies of snack foods (Ratneshwar and Shocker 1991).

However, it is also possible that people have systematically different representations and that there might not be consensus across consumers. Literature has found that representations of the same objects can vary across cultures (Lopez et al. 1997) and occupations (Medin et al. 1997). For example, the way people represent animals and trees is influenced by their environment and the functions the objects fulfill. If there are large differences between consumer segments, there could be discrepancy in how people represent expenditures. Therefore, we first explore the competing hypotheses of whether there is consensus in people's representations.

H1a: Consumers represent expenditures similarly, establishing consensus across their individual taxonomies.

H1b: Consumers do not reach consensus in how they represent expenditures—people's taxonomies are substantially different.

Also, if a hierarchical representation is unnatural to people, their taxonomy might be different across time because there could be large variances in how they choose to build a taxonomy out of expenditures. Consequently, if people form similar hierarchical representations of expenditures over time, it would be suggestive of a hierarchical representation. Similar representations over time is also indictive of some stability in the way people represent relationships between expenditures, which would make it easier for people to recruit the taxonomic distance when making spending and savings decisions.

Second, a taxonomic representation predicts how much people will adjust their spending on items at different levels of the hierarchy. Specifically, we predict that, all else equal, consumers' spending adjustment on an item will be influenced by the taxonomic distance (i.e., the level at which items are categorized together) between the item they just spent on and an item they are considering purchasing. When a consumer deviates from her budget (i.e., overspends or underspends), she would adjust spending for items that are more closely related than those distantly related. If a consumer categorizes movies and scarves together as discretionary spending and treats bread as a necessity, the taxonomic distance between scarves and movies would be closer than that between scarves and bread. Consequently, we would predict that the consumer adjusts more on movies (i.e., the closer item) than on bread (i.e., the farther item).[2]

Several restrictions apply to the proposition that hierarchical taxonomy can predict spending adjustment. First, we do not suggest or believe that taxonomic distance is the only factor that consumers consider. Characteristics of expenditures such as the necessity of the spending will influence consumers spending intentions, and there are necessities that people cannot adjust onto such as rent or phone bills (Zhang et al. 2020). We argue that controlling for the other plausible factors, taxonomic distance will influence spending behaviors on items that have room for adjustment. Further, spending deviations from necessities might not allow room for choices on other items that need to be cut back on: if someone struggles to pay rent, then they will cut back on all the expenditures they possibly can. Therefore, we follow the previous literature (Heath and Soll 1996, Cheema and Soman 2006) and focus on spending adjustments in discretionary spending—that is, purchases that are not absolute necessities. Finally, we note that

[2] We note that this hypothesis holds regardless of whether H1a or H1b obtains. If H1b obtains, we can use information about each person's taxonomy of expenditures to predict her spending and saving decisions. If H1a obtains, we can use the consensus taxonomy to make predictions about a group's decisions.

we are attempting to characterize how people adjust their spending and savings when they are thinking about alternatives. This is the standard shopping context as people are almost always evaluating alternatives on product shelves or recommended lists on webpages. Therefore, we will always present alternatives for people to evaluate in our studies.

H2: All else equal, when consumers deviate from their budget, their taxonomy of expenditures influences their spending and saving decisions. Specifically, when they deviate from their budget (i.e., over- or under-spend) on an item, they are likely to adjust more on items that are closer in taxonomic distance to the purchased item.

Finally, if people naturally represent expenditures in hierarchically-organized categories, they will spontaneously adjust their spending in accordance with taxonomic distances. Here we use the term "spontaneous" to highlight that—contrary to previous literature (Heath and Soll 1996, Cheema and Soman 2006)—we do not prompt participants with predefined categories. We predict that consumers will make spending adjustments even when not explicitly prompted with budget category labels.

H3: People spontaneously recruit their taxonomy when making spending decisions. That is, the adjustment stated in H2 happens even without reminding people of the categories to which the expenditures belong.

In what follows, we first investigate how expenditures are represented. Then, we assess whether people spontaneously adjust their spending based on their representations even when they are not primed to think about budgeting categories.

Study Overview

In seven studies, we examine how consumers represent expenditures and how their representation of expenditures influences their budgetary decisions. In study 1a, we elicit

consumers' represented taxonomy of expenditures and find consensus in their representations. We find that people's representations are relatively stable across time in Study 1b. Study 2a and 2b investigate how people make spending and savings adjustments based on both each respondent's own individual taxonomy and the aggregate taxonomy, and Study 2c finds that people's hierarchical taxonomies influence their budgeting behaviors in a way not accounted for by the principles of substitutability and complementarity. The results are consistent with the hypothesis that people adjust their spending more on items that are taxonomically closer (whether the items are substitutes, complements, both, or neither). Studies 3a and 3b use consequential choices to test whether similar patterns generalize to purchasing items on promotion. Finally, Study 4 examines more than seven million trips to grocery stores across a decade, and finds analogous patterns—consumers spend more on items when taxonomically close items are on sale. The findings from all studies suggest that mental representation is important for studying how people budget and spend. The Web Appendix and all lab study materials (pre-registration, data, and analyses) are available on OSF (https://osf.io/bcknx/?view_only=94e720c38d014813bd1f672682d5b9a2).

Study 1: Representation of Expenditures

Study 1a investigates whether there is consensus in people's representations (H1) of expenditures while study 1b investigates the consistency of the representations across 3 months. Regardless of whether consensus is obtained, we will further qualitatively examine how consumers represent expenditures.

Study 1a: Initial Exploration

Participants & Procedures. Twenty-seven participants (all located in the US) were recruited on Amazon Mechanical Turk (MTurk). The average age was 38.4 and 37.0% were

female. In this study, participants were prompted to think about expenditures by performing a successive pile-sort task (Boster 1994), which is widely used to examine people's representation of concepts (Lopez et al. 1997; Medin et al. 1997).

Participants performed the successive pile-sort on a web interface (Figure 2). Participants saw a set of 64 cards labeled with expenditures (e.g., rent, gas, movie tickets, etc.). The 64 expenditures were selected from pilot studies described in Web Appendix A, and the expenditures were presented in randomized order. Participants were instructed to group the expenditures into categories according to which expenditures they thought would go together. They could construct a group of any size or add however many categories they would like to construct groupings natural to them. After they formed the initial level of groupings, the participants were asked to review the categories they had just put together for the merging stage. They were asked to put together the categories that were most similar to each other, forming a higher level of less specific and larger categories. Then, the interface implemented the splitting stage presenting all expenditures in one initial-level group. Participants were asked to split out the items from the initial grouping that were the most different from each other into smaller (lower-level) groups. They then repeated the splitting stage for all the initial groups. Participants were familiarized with the interface and the merging and splitting task with a simple set of symbol stimuli (Web Appendix B) before they proceeded to the main task.

Analysis and Results. We first translated participants' sorting into taxonomic distances. An item had distance 0 with itself, while the items grouped in the lower-level group (i.e., the group after the splitting stage) had a distance of 1. Items grouped in the initial group but not in the lower-level group had a distance 2. Those that shared only the higher-level grouping had a distance of 3, while those never grouped together had a distance of 4. Each participant's

categorization result was characterized as a 64 by 64 distance matrix, which represented all pairwise distances between items.

Consensus Result. We then examined whether there was consensus among the participants (H1). Consensus was assessed using the Cultural Consensus Model (CCM; Romney, Weller, and Batchelder 1986, Medin et al. 1997). Each participant's distance matrix was correlated with every other participant, which yielded a 27 by 27 correlation matrix. Then, a principal component analysis was performed on the inter-subject correlation matrix. According to CCM, there is consensus if the loadings are positive and the first latent root (eigenvalue) is relatively large compared to the rest (with a rule-of-thumb being 3:1; Weller, 2007). We found that the loading for all participants was positive, and the first eigenvalue was 12.23, which is large relative to the second eigenvalue of 1.39. These results suggest that there is consensus in people's representations: people share the understanding that shampoo and sunscreen are closer together in representation than shampoo and movie tickets.

Aggregate Level Categorization. Because there was consensus, we further examined the aggregate representation with the implication that the structure and the distance between items would generally hold for most of our participants.

We first identified the initial categories that were constructed on the aggregate level. These categories reflect the general concepts that consumers have when budgeting, and they can be informative of the categorical labels that researchers and marketers might adopt. To obtain the categories, we first obtained the aggregate distance matrix by summing up all individuals' distance matrices. Then, we mapped the aggregate distance matrix onto two dimensions with multidimensional scaling (MDS; Kruskal and Wish, 1978; Shepard, 1962), which yielded the value of each item with respect to each of the two dimensions. To identify the categories, we

performed k-means clustering on the MDS values. We chose five as the number of clusters for two reasons. First, five was the mode of the number of groups that participants constructed on the initial-sort (range = 3-18). Further, five clusters allowed for a small within-cluster sum-of-squares as well as variation in the categories.

Applying five clusters to the aggregate level distance matrix, we obtained a grouping presented in figure 3. The five groups that consumers readily have in mind are roughly centered around liabilities (e.g., rent), groceries (e.g., juice), household products (e.g., shampoo), clothing items (e.g., shirts), and entertainment (e.g., dining out, airplane tickets). We conjecture that the two dimensions roughly correspond to 1) the typical spending on the item and 2) the necessity of spending on the item. Further, the dendrogram of the average item distances is presented in Figure 4[3]. Among the groups, liabilities have high ratings on necessity and items in this group often have a rigid budget (Appendix C), and hence we will focus on the rest of the groups when selecting stimuli for spending adjustment in the following studies.

Study 1b: Longitudinal Study

Participants & Procedures. Two hundred and one participants (all located in the US) were recruited on Prolific to participate in a longitudinal study. The study asked them to perform the same successive-pile sort as study 1a twice, using two waves of data collection separated by three months, to examine the stability of people's representations. The first wave of data collection was in January 2022 while the second wave was in April 2022, 13 to 15 weeks after the first wave. 146 (72.6%) participants started the second wave while 131 (65.2%) participants

[3] Note that the groupings in Figure 4 are slightly different from those in Figure 3. This discrepancy is reasonable because the clustering was based on the reduced dimensions derived from MDS, while the dendrogram was based directly on the distance matrix. The two visualizations allow us to examine the categories and the dimensions people categorize on.

completed both parts. The average age in the first wave was 33.2 and 155 (77.1%) were female. The average age in the second wave was 34.8 and 105 (80.2%) were female.

Analyses & Results. We first tested whether there is selective attrition. We ran a logistic regression using our collected demographic variables (i.e., age, income, gender) to predict whether participants finished the second wave of data collection. Other than age, which marginally predicts finishing in the second wave (β = 0.03, z = 1.93, p = .053), no other collected demographic variables significantly predicted the participation in the second wave (*p*s > .05). Further, the model predicts attrition no better than a model that simply predicts the base rate of participation in the second wave (i.e., a model with only an intercept that is not conditional on age, gender, and income) ($\chi^2(4) = 2.98$, $p = .41$). No demographic variables predicted the correlation between representations in the two waves (|*t*|'s < .65, *p*s > .5).

Next, we examine the stability of people's representation across time. The correlation between the aggregate matrices in the two waves is 0.99. The average correlation between each individual's distance matrices is 0.54 (*SD* = 0.16, 25^{Th} Percentile = .41, 75^{th} Percentile = .66). To interpret this value in context, we performed a comparison test. Each person who participated in the second wave was paired with a randomly drawn different person from the first wave and the correlation between their representations was calculated. The average correlation when one was not paired with herself is 0.42. We repeated this process of random pairing and correlation calculation 5000 times and obtained a bootstrapped 95% CI of [39.9, 43.3]. This suggests that people's representations are relatively stable over time.

Discussion

Studies 1a and 1b used a bottom-up approach to recover consumers' representations of expenditures by asking participants to successively sort various expenditures, and the results

have several theoretical and practical implications. First, the recovered taxonomy of expenditures sheds light on the categories of expenditure that most consumers might entertain. Some of the five initial level categories we recovered (liabilities, groceries, household products, clothing items and entertainment spending) are similar to those used in previous research on mental budgeting (e.g., entertainment, food, and clothing, Heath and Soll 1996). Practically, if budgeting apps structure their categories in a way that better aligns with people's representations, people might be able to better track their spending and set their budgets more accurately.

Further, study 1a also provides insight into how these categories relate to each other and how items relate to categories. For example, the category of grocery foods is close to household products but far from entertainment, and suitcase is at the boundary of the "clothing" category and the "entertainment" category. These results provide important insights for interface designs, as the design of directories and menus can benefit from understanding how expenditures are structured in people's minds. For example, people might have an easier time navigating through directories if they are more aligned with how people represent products.

In addition, we found that consumers have a relatively stable represented taxonomy over time (study 1b), and they reach a consensus in their representation of expenditures (consistent with H1a). They share the understanding that shampoo and toilet paper are closely related to each other, while shampoo and movies are not. However, we note that consensus does not mean that there is no heterogeneity in consumers' groupings. In our studies, some consumers generated three initial level groups while others generated more than a dozen, but consensus can still be obtained as long as the distances between items across individuals are highly correlated. Even though participants varied in the sizes and number of categories they created, there was consensus on the relative taxonomic distances between expenditures according to CCM.

We proposed that budgetary restraint is based, at least in part, on consumers' taxonomic distances, and recovering the representation allows us to use these empirical distances to examine consumers' spending adjustment. Specifically, we predict that when consumers deviate from their budget on a focal item, they are likely to adjust more on expenditures that are taxonomically closer rather than taxonomically distant. Because the representation of expenditures reached consensus, we can test our prediction on both the individual level (which we could also have done if consensus had not obtained—H1b) and the aggregate level. That is, how individuals represent expenditures can predict their individual spending adjustment, and the average distances between expenditures can predict the average adjustment pattern as well. In the following studies, study 2a examines individual spending and saving decisions. Study 2b examines aggregate spending adjustments while study 2c evaluates the alternative accounts of substitutability and complementarity. Studies 3 and 4 extend the domain and examine whether people's taxonomy can predict a variety of spending and saving behavior such as promotions application and grocery purchases.

Study 2: Taxonomic Distance and Spending Adjustment

Studies 2a to 2c examine the implication of a hierarchical taxonomy of expenditures on people's spending adjustment. Specifically, we propose that after overspending or underspending on an item, people will adjust more on items that are taxonomically close. We adopt our paradigm from Heath and Soll (1996) by prompting consumers to consider a spending scenario on an item (i.e., a focal item). We then elicited their adjustment on several comparison items that have different taxonomic distances with the focal item.

Study 2a: Customized Stimuli

In Study 2a, we examine if one adjusts differently for items of different taxonomic distances using each participant's individual taxonomic distances. In order to customize our stimuli to each participant's own taxonomy, we designed a two-wave survey. In the first wave, we asked participants to perform the successive pile-sort as in study 1, and we then customized the stimuli used in the second wave for every participant.

Participants and Procedures. Study 2a was a two-wave survey. We pre-registered to recruit 200 participants for the first wave and only invited back participants who were eligible for the second wave. Specifically, only participants who established a clear four-level hierarchy would be eligible for the second wave. For the second wave, we also pre-registered that we would only include participants who passed attention checks. One hundred and ninety-eight Prolific participants (all located in the US) answered the first survey. Out of the 198 people, 171 (86%) people were eligible for the second wave, and 168 answered the second survey. 161 participants passed attention checks and were included for analysis. The average age in the sample was 31.4 (range:18-77), and 57.9% were female.

The first wave of Study 2a used the same categorization task as Study 1. Participants sorted the items into categories (initial level), merged the categories (higher level) and split the initial level categories into smaller categories (lower level). Two days after the first wave of data collection, participants who were eligible were invited back for the second wave.

In the second wave, we used a 2 (spending condition: overspend vs underspend on the focal item) by 2 (scenarios) within-subject design. In each scenario, participants read that they had spent on a focal item, and they could adjust their spending on four comparison items: a lower-level item, an initial-level item, a higher-level item, and a different-category item. A lower-level item is one that was categorized together by the participant with the focal item at the

lower-level in the first wave (taxonomic distance = 1). An initial-level item is one that was categorized with the focal item at the initial level but not the lower level (distance = 2), and so forth for the higher-level item (distance = 3). The different-category item was never categorized together with the focal item at any level (distance = 4).

The focal items were chosen from an ordered list: "dining out", "shirts", "microwave", "coffee", which were close to the center of four different clusters in figure 3[4]. Here, four focal items are chosen even though we recovered five groups in study 1 because we wanted to avoid absolute necessities (the fifth group) where there might be no choice in spending adjustment. Since each participant saw two scenarios, we used the first two items for which the participant generated a four-level hierarchy (i.e., there were items categorized at each level with the focal items). For each focal item, we then selected the four comparison items that are maximally comparable with each other. To do this, we first generated all possible focal-comparison combinations of the four distances for each participant. Then we selected the one combination that minimized the difference along additional norming dimensions. These norming dimensions are collected with a separate group of participants and included questions such as how hedonic it is to spend on the item and how much one budgets for an item (see Web Appendix C for a full list of dimensions). Participants were not invited for the second wave if none of the focal items had a four-level hierarchy (e.g., if they failed to split out an initial category).

More specifically, participants first entered a price that they usually budget for the focal item. Then, they read about a scenario where they underspent [overspent] 30% on the item with

[4] The order corresponds to how spread out the clusters are in the MDS. “Dining out” is centered at the most spread-out cluster while “coffee” is centered at the least. Larger spread in the cluster indicates that the categorization is more varied, which will result in more differences in the comparison items chosen across individuals, and therefore we wanted to choose them more often.

respect to their budget. They then rated their likelihood to adjust their spending on the four comparison items (lower, initial, higher-level, and different-category item) on a 7-point scale (1-Underconsume a lot, 4 – No change, 7 – Overconsume a lot). Each participant rated their adjustments for a scenario where they overspent on the focal and another where they underspent on the focal, and the order of the scenarios was randomized. After this task, participants completed a tightwad-spendthrift measure (Rick, Cryder and Loewenstein 2008) and a propensity to plan measure (money subscale; Lynch et al. 2010) for additional controls. They then entered their age, gender and income, and were compensated for the study.

Analysis and Results. First, we assessed whether participants' sorts reached a consensus. For all the participants (N = 198) who finished the first wave (i.e., the categorization task), the first eigenvalue of the correlation matrix was considerably larger than the second eigenvalue (82.08 vs 8.50), and all the loadings were positive. These results suggest that there is consensus in people's representation of these expenditures (H1a), replicating the finding of Study 1. And again, if they had not (H1b), we could still make individual-level predictions about spending and saving decisions.

We then test whether people are more likely to adjust on items of closer distances when they deviate from the budget they set for a focal item (H2). In this case, when people overspent on the focal item, they should save more money on the items that are taxonomically close than those that are taxonomically distant. On the other hand, when people underspent on the focal item, consumers should increase their spending on the close items the most and the far items the least.

To test the effect of taxonomic distance on the spending adjustments resulting from budget deviations, we ran two regressions, one for overspending and one for underspending. We

regressed consumers' spending adjustment onto the contrast-coded comparison items' taxonomic distance (1 = lower-level; 2= initial-level; 3 = higher-level; 4 = different category) while clustering the standard error by participants. Consistent with H2, when people overspent on a focal item, they adjusted their spending less for items at greater taxonomic distances ($\beta = .10$, $t(157) = 3.57$, $p < 0.001$). However, taxonomic distance did not predict upward adjustment when there was underspending on a focal item ($\beta = .0031$, $t(157) = .13$, $p = .89$; Figure 5). The magnitude of overall adjustment was also smaller in the underspending condition than in the overspending condition (0.07 vs 0.78, $F(1,160) = 46.64$, $p < 0.001$), which is consistent with previous literature that there is an asymmetry in self-reported adjustment across events of overspending and underspending. Specifically, people report greater intentions to reduce spending after overspending, relative to their intentions to increase spending after underspending (Zhang et al. 2020).

In Web Appendix E, we report the regression models with additional controls. We included individual differences measures and norming controls (e.g., how hedonic it is to spend on the item, how much one wants to budget for the item, etc. Web Appendix C). Web Appendix E also included different specifications of the models for study 2a, such as models including focal item fixed effects and individual random effects.

Consistent with H2, when people overspent on an item, they would choose to save more money on expenditures that are closer in taxonomic distance. While most previous mental budgeting papers explicitly prompt people with category labels and ask about spending adjustments, study 2a avoids the explicit prompt. But it is possible that study 2a heightened the accessibility of these categories, even though not explicitly presented, because all participants completed the categorization task before the spending adjustment task. So, in study 2b, we

counterbalance the order of the categorization and spending-adjustment task. Further, because both study 1 and study 2a found consensus in participants' sorting of expenditures, we use stimuli selected from the aggregate taxonomy in study 1 for the spending-adjustment task in study 2b.

Study 2b: Aggregate Stimuli

Participants and Procedures. We pre-registered to recruit 400 participants and only included those who passed attention checks for analysis. Three hundred and seventy-two MTurk participants (all located in the US) participated in the survey and passed the attention check. The average age in the sample was 33.4 (range:18-77), and 51.6% were female.

Study 2b used a 2 (spending condition: overspend vs underspend on the focal item) by 5 (scenarios) within-subject design. Participants read about five purchasing scenarios on five different focal items. For each scenario, participants read that they underspent [overspent] on the item with respect to the budget. The budget and the respective deviation amount also varied across all five scenarios for more generalizable results.

Participants then rated their likelihood to adjust their spending on three comparison items on a 7-point scale (1- Underconsume a lot, 4 – No change, 7 – Overconsume a lot). The three comparison items were a close item, an intermediate item, and a far item. A close item is one that was categorized together with the focal item at the initial level by more than 90% of the participants in Study 1. An intermediate item is one that was categorized together with the focal item at the initial level by 50% of the participants. A far item was never categorized together at the initial level with the focal item. The comparison items were matched on two additional ratings collected through a norming study: (i) the likelihood of purchasing the item and (ii) the amount usually spent on the item (Web Appendix C). We included these measures because these

considerations may influence people's intentions to adjust spending on a given item. The specific sets of stimuli are presented in table 1.

Participants also finished the same successive pile-sort task as study 1 in addition to the spending-adjustment task. The order in which they completed the pile-sort task and the spending-adjustment task was counterbalanced across participants. They then entered their age, gender, and income, and were compensated for the study.

Analysis and Results. We first check whether the stimuli we selected from study 1 map onto participants' taxonomy in study 2b. In other words, we check whether this group of participants categorized the close item with the focal item most of the time, intermediate item half of the time, and far item almost never. Averaging across five scenarios, we found that the focal items were categorized with the same-category items 83.2% of the time, the marginal-category items 54.8% of the time, and the different-category items 12.8% of the time. The pattern was qualitatively the same for all five sets. This suggests that the taxonomic relationship between stimuli is consistent with study 1.

To test the effect of different levels in the hierarchy on spending adjustment (H2), we ran two regressions, one for overspending and one for underspending. We regressed consumers' spending adjustment onto the contrast-coded comparison items' taxonomic distance (-1 = close; 0= intermediate; 1 = far), order of the tasks (categorization-first = 0.5, rating-first = -0.5), and the interaction between the two, while clustering the standard error by participants. When people overspent on the focal item, they reduced their consumption more for items that are more likely to be grouped together with the focal item ($\beta = 0.14$, $t(360) = 5.99$, $p < .001$; Figure 6). This result conceptually replicates the past findings (Heath and Soll 1996) that people restrict their use of money for purchases within the same category. Similarly, when people underspent on their

focal item, they adjusted consumption upward for items that were more likely to be grouped together with the focal item (β = -0.08, $t(358)$ = -4.24, p <.001). The magnitude of adjustment in the underspending condition is smaller than those in the overspending condition (0.03 vs 1.17, $F(1,353)$ = 137.4, p < 0.001), which is consistent with study 2a. Participants compensated for spending less or more than usual on focal items by spending or saving more on the taxonomically-close items.

These results are unaffected by the order of tasks (main effects and interactions, $|t|$s < 1, ps >0.15). Further, they remained significant when we controlled for additional norming measures and individual random effects (Web Appendix E). We also report results using different ways of calculating distances including the distances from the multidimensional scaling and hierarchical clustering. The results remained qualitatively the same across changes in model specifications, controls and ways of computing distances (Web Appendix E).

We further investigated whether people spontaneously adjusted their spending based on the taxonomic distance even in absence of the taxonomy-elicitation task. Specifically, we repeated the same analyses on the group that completed the spending adjustment task before the categorization task. As indicated by the lack of order effects reported above, the pattern was the same: people adjusted more when the comparison items were taxonomically closer to the focal item (Overspending: β =.12 , $t(192)$ = 3.75, p < .001; Underspending: = -.08, $t(192)$ = -2.76, p = 0.003). In other words, even when people are not explicitly prompted to think of a taxonomy of expenditures, taxonomic distance relates to people's spending and saving decisions.

Because we also collected each individual participant's pile-sort, we further verified whether an individual's own taxonomy predicts her adjustment pattern. We recoded each comparison item with the taxonomic distance from each participant's own sorting (1 = lower-

level; 2 = initial-level; 3 = higher-level; 4 = different category). Then, we performed the two regressions, predicting spending adjustment in overspending and underspending with distances from individualized taxonomy while controlling for the additional norming measures we collected. The result is the same: participants adjusted their spending more for taxonomically close items (Overspending: $\beta = 0.05$, $t(357) = 2.61$, $p = .005$; Underspending: = -0.04, $t(355) = -1.89$, $p = 0.058$).

Study 2c: Difference from Substitutes and Complements

Study 2c examines the possibility that taxonomic distance is simply capturing substitutive relationships between products, as products that substitute for each other produce positive cross-product elasticities, like those we observed in Study 2a and 2b. We theorize that taxonomic distance is distinct from subsitutability, and Study 2c assesses this claim empirically. Study 2c uses a similar design as Study 2b with a larger set of stimuli (24 focal products total), and controls for substitutability and complementarity ratings across product pairs.

Participants and Procedures. We pre-registered to recruit 400 participants and only included those who passed attention checks for analysis. Three hundred and seventy-six Prolific participants (all located in the US) participated in the survey and passed the attention check. The average age in the sample was 37.88 (range:18-80), and 48.4% were female.

Prior to running study 2c, we conducted a norming study to collect the substitutability ratings and complementarity ratings for pairs of products that are often grouped together that the lowest level (Web Appendix D). Then, we selected three unique product pairs that are very high on rated substitutability, three that are very low on rated substitutability, three that are very high on rated complementarity, and three that are very low on rated complementarity, resulting in a total of 12 product pairs. We used all the items in the 12 pairs as the focal item for study 2c,

resulting in a total of 24 focal items (i.e., spending scenarios). This selection aimed to maximize the variance in the substitutability and complementarity ratings among the products that are often grouped together at the lowest level, so that we can best capture the effect of these cross-product relationships. Among the products we selected, there is a negative correlation between taxonomic distance and substitutability ratings (Pearson's $r = -0.72$, $p < .001$) and a negative correlation between taxonomic distance and complementarity ratings (Pearson's $r = -0.38$, $p = .002$). In other words, taxonomic distance is related not only to how much a pair of purchases substitute for each other, but also to how much these purchases complement each other. So, variation along taxonomic distances does not simply equate to variance in substitutability patterns.

Study 2c used a 2 (spending condition: overspend vs underspend on the focal item) by 24 (scenarios) mixed design. Each participant read a set of six purchasing scenarios, randomly drawn from four sets (24 scenarios total) on six different focal items. We predefined the sets such that all the focal items in a set were relatively different from each other.

Each participant read that they either over or underspent on all the focal products: "You realized that the [focal product] you bought last week was a lot more expensive (a lot cheaper) than you thought it was." Similar to study 2b, participants then rated their likelihood to adjust their spending on three comparison items on a 7-point scale (1- Underconsume a lot, 4 – No change, 7 – Overconsume a lot). The three comparison items were a close item, an intermediate item, and a far item.

Analysis and Results. For each spending condition, we preregistered to run two regressions. One predicts the likelihood-to-adjust ratings with 1) the taxonomic distance between the focal and 2) how substitutive the target and the focal is to each other, controlling for several

other dimensions of norming data that we collected such as how hedonic the spending was and how much people wanted to control the spending on the item. The other regression is the same other than replacing the substitutability rating with the complementarity rating. Taxonomic distance was significant in all four regressions in the direction expected. Specifically, when participants overspent on the focal, they adjusted down more on items closer in distance (Controlling for substitutability: $\beta = 0.167$, $t(176) = 4.68$, $p < .001$; Controlling for complementarity $\beta = 0.173$, $t(176) = 7.86$, $p < .001$; Controlling for both $\beta = 0.152$, $t(175) = 4.25$, $p < .001$). When they underspent on the focal, they adjusted up more on items closer in distance (Controlling for substitutability: $\beta = -0.107$, $t(188) = -3.60$, $p < .001$; Controlling for complementarity $\beta = -0.081$, $t(188) = -4.32$, $p < .001$; Controlling for both $\beta = -0.112$, $t(187) = -3.73$, $p < .001$). Full regressions are available in Web Appendix E. The raw ratings and the esimtated marginal mean ratings after controlling for both substitutability and complementarity are further shown in Figure 7. These results suggest that even when the experiment is set up to detect the effects of substitutability and complementarity, we find that taxonomic distance provides an additional explanation in people's behaviors.

Discussion

Consistent with H2, studies 2a-c find that after people spent on an item, they would intend to adjust more on expenditures that are closer in representation. This result holds regardless of whether we use stimuli selected from a participant's individual taxonomy or from an aggregate taxonomy, whether people complete the categorization task before making spending adjustments, and whether we control for substitute and complementary relationships (study 2c). These findings are consistent with the notion that people reference a taxonomy of

expenditures when making decisions about spending on alternatives, and the findings are not explained by the substitutability and complementary relationship between the expenditures.

The results of our studies extended the budgeting literature in several ways. First, in these studies, the categories participants reference were not prompted in the immediate context of the task. This suggests that the categories of expenditures we recover align with how consumers represent and restrict their use of money. Further, our studies provide further insights on budgeting behavior beyond dichotomous categorical relationships of budgets. People make differential adjustments on spending depending on the level that the items are categorized together (study 2a). Finally, all studies used deviation from budget restraint in both directions (i.e., overspending and underspending) and found patterns of adjustment in both directions. This finding further suggests that both spending and saving decisions can be predicted by people's taxonomies.

Study 2 examined the connection between people's represented taxonomy and their spending adjustment intentions, but all studies still asked participants to categorize the expenditures[5]. In studies 3 and 4, we test H3—whether participants naturally adjust their spending differentially depending on the taxonomic distance across a variety of spending decisions without ever doing the categorization task. In addition, since we found relatively limited increases in spending intentions in studies 2a and 2b when people saved money on the focal, we further test whether taxonomic distance can lead to upward adjustment on spending. Therefore, we focus our investigation on behaviors related to spending adjustments when there is underspending on the focal in both studies 3 and 4.

Study 3: Taxonomic Distance and Promotions

[5] Although for half of Study 2b's participants, the categorization task followed the decision task.

Studies 3a, and 3b examine consumers' savings adjustment behavior across a variety of spending decisions. In these studies, participants read that they purchased a focal item, and the focal item came with a discount that they could apply to a comparison item. Participants in study 3a ranked the comparison items that they wanted to apply a promotion to, and participants in study 3b made an incentivized choice of a comparison item that they could get for free. We hypothesize that a product purchase would prime the more specific budgeting categories associated with the purchase. Consequently, the purchase would more strongly influence the saving decisions of a close item than a far item, and people would be more likely to apply promotions to a close item.

Study 3a: Interaction with Typicality

Participants and Procedures. Four hundred and two MTurk participants (all located in the US) responded to the survey while 356 passed the attention check. The average age in the sample was 40.6 (range:18-73), and 44.1% were female.

Participants assumed that they were purchasing all items on a shopping list of four items. They learned that one of the items (the focal item) that they purchased comes with a promotion—a price discount to be applied to another item on the list. Then, they ranked the items by how much they wanted to apply the promotion to each item (1 - I want to apply the promotion to this item the most; 3 - I want to apply the promotion to this item the least). We expected most participants to apply the promotions to the comparison item that is the closest in their representation of these expenditures.

Study 3a used a 2 (promotion magnitude: large – 40% off vs small – 10% off) by 5 (scenarios) by 2 (focal typicality: typical vs atypical) within subject design. Each scenario consisted of a focal item that participants had already purchased and three comparison items that

they could apply a discount promotion to: a close item, an intermediate item, or a far item. Studies 3a and 3b used a new set of stimuli that consists of 50 products that people can purchase in large wholesale stores, and a norming study established the aggregate distance between the items (Web Appendix F). We included focal items of varying typicality in order to explore whether the typical focal items could prime the categories better and thereby induce a stronger effect (Rosch and Mervis 1975; Heath and Soll 1996; Osherson et al. 1990). We collected typicality ratings in a separate norming study, where participants rated the typicality of expenditures with respect to the initial categories they constructed in a successive pile-sort task (Web Appendix G).[6] Each participant saw five scenarios, and the promotion magnitude and the focal typicality were randomized for each scenario.

Analysis and Results. To test the effect of taxonomic distance on people's preferences for applying the promotions, we ran an ordinal logistic regression (McCullagh 1980). We regressed consumers' rankings of the items onto the items' aggregate distance from the focal (centered at 0), contrast-coded magnitude (0.5 = 40% off; -0.5 = 10% off), contrast-coded typicality (0.5 = typical focal; -0.5 = atypical focal), and the interaction between distance and magnitude as well as distance and typicality. Consistent with our hypothesis, people preferred to apply promotions to taxonomically close comparison items: there was a positive effect of distance on ranking (β = .17, Wald's Z = 6.56, $p < .001$), and more closest comparison items were ranked the highest (42.5%, vs 26.8% (mid distance) and 30.6% (far distance), $\chi^2(2) = 106.19$, $p < .001$, figure 7). We also observed a marginally significant tendency such that when the focal item is more typical, the effect of taxonomic distance is more prominent (interaction between distance and

[6] Here again, we did not prompt participants with categories, but due to the high level of agreement in categorization, some items emerged as more typical than others on average.

typicality β = .09, Wald's Z = 1.67, p = .09). The effects were consistent regardless of the promotion magnitude (interaction between distance and magnitude β = -.04, Wald's Z = -0.78, p = .43, ns). These patterns persist when we include additional controls such as how substitutable the items are for each other (see Web Appendix H).

Study 3b: Choosing Free Product

Whereas study 3a asked participants to rank comparison items, study 3b tested a different dependent variable: choice. Participants chose one comparison item that they would like to get for free when the focal item was on sale, and we used a random lottery incentive where participants had a chance to realize their choices.

Participants and Procedures. We pre-registered that we would collect 300 participants. Three hundred and seven MTurk participants (all located in the US) responded to the survey and 302 people passed the attention check. The average age in the sample was 38.8 (range:20-70), and 51.7% were female.

Participants were randomly assigned to two conditions: they assumed that they bought one of two focal items (coffee or toilet paper; forced-focal condition), or chose between the two items to buy (choose-focal condition). Participants read that they spent $15 when purchasing the focal product, which was selected to be around the average price consumers usually spend on these products (Web Appendix G). Then, they learned that the focal item that they purchased comes with a buy-one-get-one-free promotion. They could apply the promotion to one of the three comparison items: detergents, pizza, and stationery products, all of which were priced around $15. The three comparison items are held constant in the study, but the distance with the focal item varied as revealed in our norming study (Web Appendix F). When the focal item is toilet paper, detergents share the closest distance, pizza shares the farthest distance, and

stationery products are in the middle. When the focal item is coffee, pizza shares the closest distance, stationery products share the farthest distance, and detergents are in the middle.

After participants made their choices, they could opt in to provide a link to select a comparison product that they usually purchase. Participants then answered how much they needed each of the comparison products in their daily life (1 – Don't need this at all; 7 – Need this a lot), and entered their age, gender and income to complete the study. After the study was complete, three participants were selected at random and given a $15 MTurk bonus with a message that reminded them of their product choice (and the product link if provided).

Analysis and Results. Across two conditions, the proportion of choosing the closest item was significantly larger than the mid-distance item (54.6% vs 16.6%, $X^2(1) = 93.86$, $p < 0.001$, $\phi_{cramer} = 0.53$; figure 8), and the far-distance item (54.6% vs 28.8%, $X^2(1) = 40.37$, $p < 0.001$, $\phi_{cramer} = 0.31$). This is consistent with the notion that when people spent on a focal item, they wanted to save on the closest item. Unexpectedly, we also observe that people chose the far-distance item more often than the mid-distance item ($X^2(1) = 12.23$, $p < 0.001$, $\phi_{cramer} = 0.27$). The pattern did not differ between the forced-focal condition and the choose-focal condition ($X^2(1) = 1.75$, $p = 0.42$, $\phi_{cramer} = 0.08$).

Discussion

Consistent with H2, studies 3a and 3b found that in the context of promotions, consumers make savings decisions according to the taxonomic distances between items. Specifically, when consumers bought a focal item that comes with a budget relief, they were more likely to apply discounts to other purchases that are taxonomically close (study 3a) and more likely to get taxonomically close items for free (study 3b). The pattern emerged for small and large discounts of the focal item (study 3a) and when people assumed or chose the focal item (study 3b). This

pattern suggests against the simple substitution account—if the item closest in taxonomic distance is just the closest substitute, people would not choose it after having just purchased the focal item. Further, consistent with H3, consumers were able to spontaneously make this adjustment: all participants in study 3 responded to the surveys without performing the categorization task.

The findings of study 3 further provide a moderator of our effect. Specifically, study 3a shows that spending adjustment was more pronounced when the focal item is typical, which suggests that deviations from those typical expenditures better prompt people to adjust based on taxonomic distance. This finding also conceptually replicates the finding from Heath and Soll (1996), which strengthens our theorization that budgeting behaviors are grounded in categorization principles and reinforces the current literature with consistent findings.

Our results also hint that the predictive power of taxonomic distance could diminish as the distance between the items gets larger. Specifically, all studies of study 3 found that when across items with mid-distance and far-distance, the response pattern became noisier and less predicted by the taxonomic distance. This could be related to "the basic level" concept investigated in the categorization literation. The aggregated mid-distance item could be the closest to the natural, basic level that people group their purchases (Murphy, 2002; Rosch et al., 1976, Collins and Quillian 1969), and items of closer distance might share subordinate level groups while those of further distance share superordinate groups. Spending on an item may be associated with the spending on a subordinate level category more easily than a superordinate level category. When the taxonomic distance is larger than the basic level people usually adopt, people might evaluate spending relative to a comprehensive mental account like total wealth (Morewedge, Holtzman, and Epley 2007) and think less about the taxonomic distance.

Our findings have several implications. First, marketers can leverage our findings to design product bundles and promotions that are more appealing to consumers. For example, a discounted bundle of toilet paper and detergents might be more appealing than a bundle of detergents and instant coffee. Similarly, marketers can improve the appeal of promotions by promoting taxonomically close items. When a promotion is contingent on purchasing another good, consumers might be more willing to claim the deal when the two items are taxonomically close to each other. Further, marketers can learn about consumers' mental representations of expenditures and design their research with an approach similar to the successive pile-sort in order to recover the consumers' representation between products.

Study 4: Examination of Grocery Purchase

In Study 4, we further examine H3—whether consumers spontaneously make savings adjustments based on products' taxonomic distances in shopping environments. Specifically, we investigate how spending on comparison items changes when focal items of different taxonomic distance is on sale . We study consumers' grocery purchase data from a decade's worth of shopping trips (~7.3 million; Consumer Panel Data[7]) collected and maintained by the Nielsen Datasets. Specifically, we expect two patterns of behavior. First, we expect consumers to increase their spending on comparison items when they purchase a given focal item on sale because we expect people to increase their spending when there is budget relief. Second, we expect the adjustment on comparison items to be larger when the focal item on sale is taxonomically close to the comparison item.

[7] Researchers' own analyses calculated (or derived) based in part on data from Nielsen Consumer LLC and marketing databases provided through the NielsenIQ Datasets at the Kilts Center for Marketing Data Center at The University of Chicago Booth School of Business. The conclusions drawn from the NielsenIQ data are those of the researchers and do not reflect the views of NielsenIQ. NielsenIQ is not responsible for, had no role in, and was not involved in analyzing and preparing the results reported herein.

Data and Stimuli Selection

The Consumer Panel Data contains a longitudinal panel (from 2004 to date) of approximately 40,000-60,000 U.S. households who continually provide information about each of the products that they buy. The dataset contained information about the shopping trips, such as the price spent and quantity bought for each product, and whether each purchased product was on sale. It also contained information about each purchasing household such as income, household size, and the number of children in the household (Kilts Nielsen Consumer Panel Dataset Manual 2020). We analyzed a decade's worth of 7.3 million purchase trips, from 2009 to 2018, examining the effect of focal item's deal on the price of comparison item purchased.

To examine spending adjustment on products of different taxonomic distances, we structured the data as follows. First, we collapsed across all sizes and brands of the same type of product and only recorded the product type[8]. For instance, Tide Free and Gentle detergents 92 oz. and Arm & Hammer 122 oz. would both be recorded as detergents. This is because we only recovered the participants' taxonomies of the expenditures, but not specific products. If one of the brands or sizes was on sale, we marked the expenditure item as on sale. We then take the sum of the price purchased for the products as the total amount spent on the expenditure item.

Second, we selected the focal and comparison product sets presented in Table 3. The products were selected from the same set of expenditures that we used to recover distances for studies 3a and 3b (Web Appendix F). Specifically, we selected items that were comparable in price and not rated as substitutes or complements with each other based on the norming study we collected (Web Appendix G). Further, we chose items that are unlikely to appear nearby each other (e.g., in the same aisle or on the same shelf) in a retail space.

[8] The type of product is given in the dataset as the "product module".

Finally, we selected households that have purchased the comparison item together with the focal item at least once in that calendar year. We attempt to investigate households that have budgets for the focal and comparison items we chose. Therefore, the items should not be exceptional (Sussman and Alter 2012) or one-time purchases for the household.

Model and Analysis

In order to test the effect of the focal item's deal on the price of comparison product purchased, we estimated the two models, one for the close focal and one for the far focal. Both of the models have the structure as the following:

$$\ln(Y_{CompPurcahsed}) = \beta_0 + \beta_{deal}X_{focalDeal} + \beta_{basket}Z + \beta_{household}\gamma + \beta_{compfixed}C$$

ln ($Y_{CompPurcahsed}$) is the log of the total spending on the comparison product. $X_{focalDeal}$ is an indicator variable of whether the focal product is on sale, and the two models would yield estimated coefficients for both the close focal (β_{deal_close}) and the far focal (β_{deal_far}). Z is a vector of control variables for each trip. It includes the total price and quantity of focal item purchased, the total quantity of the basket size, the number of product types in the basket, and the total price of the basket. γ is a vector of household characteristics including income level, household size, type of residence, and race. C indicates a fixed effect for each comparison item.

Our hypotheses generate two predictions. First, consumers would spontaneously increase their spending on the comparison items when a focal item is on sale. This implies that the β_{deal} coefficients should be positive for both close focals and far focals. Even though the expenditures vary in taxonomic distance, they are all expenditures incurred in one grocery trip, which is a higher-level category recovered in study 1. Therefore, consumers would perceive that they have budgetary resources left for the grocery trip when an item is on sale and spend it on other items in the trip. Second, we predict that the effect should be greater when the focal item is

taxonomically close to the comparison product, which means β_{deal_close} should be greater than β_{deal_far}.

Consistent with our predictions, we found that across four sets, people do increase spending on the comparison products both when taxonomically close items and taxonomically far items are on sale ($\beta_{deal\_close} = .22, t(3{,}823{,}102) = 53.61, p < .001$, and $\beta_{deal\_far} = .14, t(3{,}477{,}462) = 37.18, p < .001$). In addition, people increase spending more on the comparison product when the item on sale is taxonomically close (i.e., β_{deal_close} is significantly greater than β_{deal_far}, $z = 58.91$, $p < 0.001$). The pattern remains the same when the model controls for fixed effects of each year (Figure 9).

Discussion

Study 4 applied our findings to consumers' actual purchase behaviors, and consistent with H3, found that when shopping for everyday products, consumers' spending and savings decisions are related to taxonomic distances between items. This finding further implies that consumers' representations of expenditures predict their spending across a wide range of behaviors and across different samples of the population.

Further, our findings suggest that the representation of expenditures also provides insight into the demand relationships between products. Specifically, Study 4 tested our predictions about how the price change on one product influences the demand for other products. If the price drops on a given product, we can expect the demand on a taxonomically close item to increase. Taxonomic distances therefore might relate to, for example, cross-product elasticities, in addition to more widely studied factors like shelf placement and other marketing mix variables (Bezawada et al. 2009).

One limitation of the current finding is that we do not have information on the spatial proximity (i.e., how close two items are) of the items in the physical stores. If the focal and the comparison items appeared next to each other, it is somewhat plausible that receiving a deal on one would cause splurging on another. However, we think spatial proximity has limited influence on our result in study 4. Whereas all of the detergents are likely to be placed together, the products we chose (e.g., coffee, toilet paper) are likely to vary in their aisle placement across retailers. In addition, we chose products that we suspect are not right next to each other in the store (e.g., pizza is usually in the freezer and not next to coffee or toilet paper). In any case, the results of this study need to be evaluated together with the experimental results as they relate to our hypotheses.

General Discussion

Across seven studies, we investigated how people mentally represent expenditures and how taxonomic distance influences people's spending and saving behavior. Study 1 found that consumers have a consensus taxonomic mental representation of the expenditures used in these studies. The distances recovered from this taxonomy are related to consumers' spending and saving decisions. Specifically, after spending on a given item, people adjust their consumption more on items that are closer in distance to that item (Studies 2a, 2b and 2c). This adjustment appears to be spontaneous. People were only presented with the spending alternatives and there was no prompt of the budgeting category, but people still make adjustments based on the taxonomic distances for various spending activities such as promotion applications (Study 3a and 3b) and grocery purchases (Study 4). The results are consistent with our proposal that people represent expenditures hierarchically and use those representations in their spending decisions.

Our studies also provide preliminary insights on when people's taxonomies are likely to be more strongly versus weakly related to their spending adjustments. For example, study 3a finds that when the focal item is typical (and is thus more likely to prompt the categorical structure) people are more likely to adjust according to taxonomic distance. In addition, the predictive power of taxonomic distance seems to decrease as distance increases. If two items are very unrelated and the taxonomic distance between them is large, the category that includes both might be inaccessible or so expansive as to be best characterized as a "comprehensive" account. Taxonomic distance in these cases may not predict spending well. Specifically, taxonomic distance seems to have limited predictive power when the groupings are larger than people's initial grouping (study 3): we found that people did not always adjust the mid-distance item more than the far-distance item.

Limitations and Future Directions

We found that taxonomic distances predicted people's spending and savings decisions across multiple domains, but there are several questions left to explore. The framework we use for this paper primarily concerns what is likely a relatively stable taxonomy (Study 1b) that consumers recruit when they make spending adjustments. We think that a stable taxonomy is somewhat reasonable given that consumers often receive similarly structured shopping environments (e.g., groceries are separated into similar aisles). However, it is also likely that consumers construct topical accounts in an ad-hoc fashion given the context (Barsalou 1983, 1985; Sloman 1998; Sussman and Alter 2012). For example, ride share fare and bill for a meal might not be close in one's taxonomy, but they could be temporally close and relate to a common goal of dining out. Therefore, experiencing an unexpectedly high ride fare might lead to ordering a less expensive meal. This is an important phenomenon in mental accounting that lies beyond

the scope of the current paper. We predict that holding temporal proximity and goal-relevance constant, people would likely adjust more on expenditures that are taxonomically close, but future work is needed to examine in detail the factors that may temporarily or permanently change people's taxonomy.

This work also focuses on the situations where people are in the shopping environments and thinking about spending adjustments among alternatives with the idea of budgeting in their minds. We highlight that in these situations, people might be thinking about the relationship between purchases. One other important step to budgeting could be the planning stage, during which people formally set a budget by pre-defining the items or the category and deciding how much money they'd allocate to each category. Future work on budgeting could also focus on the prevalence and the behavior of such planning stage.

Future work could also examine individual or group differences in sorting and the mechanisms by which taxonomic distance is made more or less important for spending decisions. For example, people who keep a rigid budget for one or more of these categories will likely recruit that category or those categories more often and consequently be more influenced by the taxonomy. In contrast, consumers might be less influenced by taxonomic distance when their comprehensive resource-based account is chronically accessible, like those struggling to pay their rent and utilities. Relatedly, even though we observed consensus in people's pile sorts, there could still be individual differences in how people organize the expenditures, and future studies can examine if people differ systematically in the way they organize expenditures.

Finally, we presume that consumers' taxonomies of expenditures are multiply determined and influenced by many sources. Past research has identified a plethora of factors that affect the formation and use of categories. For example, the frequency with which two concepts co-occur

(Barsalou 1985), goals that objects relate to (Ross 1997), resemblance of features (Rosch and Mervis 1975), and location in the physical space (Solomon et al. 1999) can all influence the categories people construct. We conjecture that all of these elements play a role in how people represent expenditures. In addition, people could be influenced by knowledge specifically related to the organization of expenditures. Such knowledge would include the typical groupings consumers see in stores and websites, and the kinds of standard personal budgetary categories they have been instructed to construct and track. Future studies can investigate how the factors interact to establish people's represented taxonomy.

Theoretical Contributions

Our investigation offers several theoretical contributions. First, our studies add to the mental representation literature by characterizing how people represent expenditures. Expenditures are often goal-oriented, and categories constructed from expenditures are often *ad hoc* (e.g., "things that can be purchased at a supermarket"). Therefore, it was not obvious *ex ante* whether people represent them the same way that people represent other objects, like trees or animals. We found that similar to the way people represent natural kinds (e.g., animals and plants), people appear to represent expenditures in hierarchically organized taxonomies. People's pile sorts reflect broad consensus in how they mentally represent expenditures, and they spontaneously use this hierarchy in their spending and saving decisions. These results expand our understanding of how knowledge is organized in people's minds, which is critical for understanding consumer behavior.

Further, we identify a downstream effect of people's mental representation. Much research on mental representation investigated how people's representations influence mental processes such as the way they think (Collins and Quillian 1969), categorize (Collins and Loftus

1975), and make inferences (Randall 1976). Meanwhile, less has focused on closely connecting mental representation to people's observable behavior (cf. Atran et al. 1999). Our work helps bridge this gap by suggesting that people recruit their mental representation when they are making spending and savings decisions, which are concrete, observable behaviors. Specifically, people's mental representation of expenditures helps determine which items they choose to adjust their spending on as well as the magnitude of their spending adjustment. They adjust the most on items that are taxonomically closest to the focal item. These results in turn emphasize the importance of studying people's representations (cf. Reinholtz et al. 2015) and highlight the connection between one's cognitive processes and consumption behaviors (Bartels and Johnson 2015).

Moreover, our proposed structure contributes to the current budgeting literature by relaxing the often-used binary distinction of "within-category" and "out-of-category" (Heath and Soll 1996). Although toilet paper and coffee are not in the same lower-level category, they might be readily categorized together at a higher level as grocery purchases. In study 4, all the items are in the grocery category, yet receiving discounts on one can still impact the price purchased for another if the two items are taxonomically close. A binary distinction of expenditure categorization can therefore be restrictive and inaccurate for predicting the spending adjustment across the expenditures that consumers consider on a daily basis.

Finally, our recovered representations can connect to many existing concepts such as typical expenses, exceptional expenses, and malleable expenses. Because we mapped out relationships between expenditures and recovered categories as clusters (study 1), we can make predictions about how typical a given expenditure is for a category (Heath and Soll 1996; Reinholtz et al. 2015). Specifically, items closer to the center of a cluster should be more typical

of their categories. Similarly, items in-between two clusters could be more malleable and could be assigned to multiple categories (e.g., "dining out" as "food" or "entertainment"; Cheema and Soman 2006). For example, rent would be typical to the liability category whereas gym membership is more malleable between a liability or entertainment expenditure. Similarly, we could also identify items that are likely to be treated as exceptional expenses (Sussman and Alter 2012) as items that do not seem to be categorized with others in the dendrogram, like flowers and suitcases.

Marketing Implications

Consumers experience budgeting deviations all the time. Recently, high inflation, unemployment, and other income and price shocks push consumers to constantly adjust their spendings and savings. Our work provides some insights on how marketers can approach the questions of predicting and helping with consumers' spending adjustment. Credit card companies and personal finance apps can promote better consumption and budgeting behaviors when they better understand people's representation of expenditures. For example, if these companies want to decrease overspending (e.g., to lower default rates), they might facilitate the labeling and tracking of expenditures using categories that are natural to consumers. When the labels attached to consumers' card spending are consistent with consumers' representation, consumers might find it easier to track and budget their spending. Alternatively, they could also label expenditures at a level that is more specific to people's natural level of representation, because prompting people to consider smaller, lower-level categories can be helpful in restricting people's spending (Thaler and Shefrin 1981; Krishnamurthy and Prokopec 2010). For example, they could detail expenditures as "fast food" and "grocery food" instead of using the higher-level "food" label. By

eliciting multiple levels of expenditure grouping (figure 4; dendrogram in Study 1), we provide insights on how to design labels that are more or less specific to people's natural grouping.

In addition to probing how consumers think about expenditures, we highlight the approach of the successive pile-sort method of investigating relationships between concepts relevant to marketing. Marketers can use successive pile-sort to understand consumers' representations for brands and their products. For example, marketers can use the successive pile-sort to generate perceptual maps between brands (Hauser and Koppelman 1979) and better understand the demand relationships between them: brands organized together at low levels may likely be substitutes. Alternatively, a brand considering launching a new product (i.e., product extensions) can use the successive pile-sort method to examine whether the product fits the brand (Aaker and Keller 1990; Bottomley and Holden 2001). Specifically, the brand can use the successive pile-sort method to assess the distance between current products and the new product. This allows the brand to further gauge whether the new product should be introduced (Parker et al. 2017) and what the marketing message should be (Moreau, Markman and Lehmann 2001).

Further, taxonomies like the ones we have recovered might provide important insights on directory and menu designs. Companies and marketers that organize items for consumers can benefit from understanding how consumers naturally think about expenditures. They can better structure or position the products into a hierarchy that is consistent with consumers' representation for improved product search experience. For example, we found that consumers represent food and daily products closer than food and clothing (study 1). An online supermarket can benefit from knowing this structure, so that they can organize their search directory accordingly (i.e., putting foods and daily products closer in the search menu) because a well-organized directory can reduce search complexity (Jacko and Salvendy 1996; Miller 1981).

Similarly, if people have an easier time navigating when layouts match their representations, then marketers organizing a physical store could also place food and daily products closer to improve shoppers' experience and satisfaction.

Finally, even though participants in our studies reached a consensus in their sorting, marketers can still benefit from the cases when there is a lack of consensus in consumers' representations. When consumers' representations of products or brands diverge, it implies that people have heterogeneous associations for a given product or brand. Marketers could consider using these differences to characterize different consumer segments. For example, consumers who more readily group together bread and cheese might be systematically different from consumers who more readily group together bread and milk. Understanding how other differences in consumers relate to differences in sorting might allow marketers to better target their consumers.

The current paper develops a framework from theories of categories and concepts and applies a method of investigation from that literature to ask new questions about mental accounting and mental budgeting. The studies investigated consumers' representation of expenditures and related the recovered taxonomic distances to spending and saving decisions. We suspect the methods here could be used to investigate hypotheses beyond budgeting-related decisions. By connecting people's cognitive processes with consequential behavior, we hope the paper invites more investigations that apply methods used to identify cognitive processes in their study of consumer behavior (Bartels and Johnson 2015).

Chapter 2: Prediction by Replication: People Prefer Prediction Algorithms That Replicate the Event Being Predicted

Abstract

People can use algorithms to forecast the outcome of to-be-determined events like how long a drive will take, how well a job applicant will perform, and how much they will enjoy a movie. However, little is known about what types of algorithms people prefer to use. We propose that people prefer prediction algorithms that replicate the event being predicted by going through the same process that generates event outcomes, and we dub this property "replicativeness". For example, to predict the outcome of a die roll, people like to roll a die instead of using a different random process. In seven studies, we find that people like prediction algorithms more when they perceive those algorithms to be more replicative, even when this methodology leads to poor performance.

Managers and consumers face a multitude of scenarios with unknown outcomes and need to make predictions in order to make choices and judgments. Investment managers predict how much assets will be worth in the future, shoppers predict how much they will enjoy a product, sport betters predict which team will win a game, managers predict how well a job applicant is going to perform, and so on. To obtain such predictions, people can often choose between different prediction algorithms, which we define as decision rules, step-by-step processes, statistical models, or programs that offer predictions of the unknown outcomes.

People have more prediction algorithms at their disposal than ever before. Sport betters can consult many different prediction models to inform their bets, investment managers can choose between many models to construct their portfolios and make purchase decisions, and hiring managers can use algorithms or their intuition in their hiring processes. Further, many organizations offer prediction algorithms as their core product; expert forecasters like Nate Silver drive consumers to their websites and publications with their prediction algorithms, navigation tools like google maps use algorithms to help people predict how long a trip will take and which route is the fastest, and weather predicting almanacs (e.g., the Farmer's Almanac) are some of the longest running publications in the United States.

Although academics have begun to investigate people's choices between prediction algorithms in many domains, it is still unclear exactly how people make these choices. What is clear from the literature is that people do not simply choose whichever algorithm performs the best. For example, a large body of literature has shown that although prediction algorithms typically outperform human judgment, people often elect to use human judgment instead. This phenomenon manifests in a wide variety of consequential domains (Dawes et al. 1989; Meehl 1954; Yeomans et al. 2019; Longoni et al. 2019; Kleinberg et al. 2018), including medical

decisions (Gallagher 2017; Leachman and Merlino 2017; Lohr 2016) and judicial decisions (Kleinberg et al. 2018). Further it not only prevails among laypeople (Castelo et al. 2019; Dietvorst and Bharti 2020; Dietvorst et al. 2015; Longoni et al. 2019), but also professionals (Fildes and Goodwin 2007; Logg et al. 2019; Vrieze and Grove 2009). Thus, people may choose not to use prediction algorithms even when they outperform alternative options.

However, people's preferences are not as simple as a general distaste for prediction algorithms either. For example, people often prefer prediction algorithms to human judgment in relatively objective domains (Castelo et al. 2019), in domains with little irreducible uncertainty (Dietvorst and Bharti 2020), before getting performance feedback (Dietvorst et al. 2015), and when algorithms demonstrate an ability to improve over time (Berger et al. 2021). These results suggest that people are open to using some algorithms in some domains, but more work is needed to understand exactly what factors affect this choice.

Relatedly, past research has often treated prediction algorithms as a monolith, exposing study participants to one particular algorithm and then drawing conclusions about participants' general propensity to use any algorithm. However, there are infinite potential algorithms, and people may like some more than others. Indeed, algorithmic predictors can do anything from simple averaging to implementing complex machine learning models and many other approaches that have yet to be conceived. Therefore, to better understand people's use of algorithmic predictors, and learn how to boost the use of algorithms that can help people make more accurate predictions, it is important to investigate what types of algorithms people prefer.

In this paper, we aim to build understanding of which types of algorithms managers and laypeople are likely to adopt, and which types they may be more hesitant to use. Further, we aim to build understanding of how managers and marketers can describe prediction algorithms in

terms (and build prediction algorithms in ways) that boost adoption of these algorithms. In line with this goal, we identify a factor that people assess specifically when evaluating prediction algorithms: how replicative the prediction algorithm is of the event being predicted.

Defining Replicativeness

We define replicative prediction algorithms as those that generate predictions by reproducing the process of the event being predicted. Replicative algorithms rely on the logic that the same process leads to the same outcome, and thus, try to predict what outcome will occur by seeing what happens when as similar of a process as possible generates an outcome. For example, a replicative algorithm for predicting the duration of a car trip may simulate it – reproduce the car trip given current traffic flows and see how long it takes. In contrast, a non-replicative algorithm would use a process that is dissimilar to driving the route, such as using a simple formula to calculate the expected duration of the trip given the base rates for different types of roads. Replicative algorithms can also try to reproduce the process of the event being predicted by looking to past cases that are as similar as possible to that event. For example, a replicative algorithm for predicting the duration of a car trip could also find nearly identical past trips and report how long they took. Such an algorithm may look up past trips with the same origin, destination, time of day, weather, and so on. A non-replicative algorithm may rely on data from past trips that do not resemble the current trip.

In some domains, the processes that generate outcomes are transparent, and thus, assessing the replicativeness of a prediction algorithm may be straightforward. For example, because the distance that a projectile travels in a vacuum before striking the ground depends on its velocity, launch angle, and height, a perfectly replicative algorithm for predicting the distance of a particular launch would simply stage an identical launch and use the outcome as its

prediction. However, many prediction events have unclear processes. For example, we do not understand exactly how weather events develop or how stock market returns are generated. In these cases, people's perceptions of replicativeness will depend on their beliefs about the process that generates outcomes and their judgments of how well that process is captured by a prediction algorithm.

Replicativeness is therefore a perception people have when comparing two processes: the process that a prediction algorithm uses to make predictions and the process that the predicted event goes through to generate outcomes. On the surface, judgments of replicativeness may seem similar to judgments of representativeness, which people make when they are judging outcomes (instead of processes) (Tversky and Kahneman 1971, 1974, 1982). Kahneman and Tversky (1972) propose that outcomes are perceived to be more representative when they reflect the "essential properties" of the "process or model" that produces them (Tversky and Kahneman 1982, p.85). For example, people often judge sequences of coin flips that alternate between heads and tails (the outcome) to be more likely than those that contain long runs because alternating sequences reflect the odds (50% heads & 50% tails) and randomness of generation process (i.e., flipping a coin).

In contrast, when judging replicativeness, people compare two processes (a process that makes predictions and the process that generates outcomes), instead of judging the likelihood of an outcome (as is the case for representativeness). Thus, while past literature has established that people often use representativeness to judge outcomes, we propose that people use replicativeness to judge prediction algorithms. Returning to the example of coin flips, people would likely judge flipping a coin as a more replicative way of predicting coin flips than drawing

bingo balls that say either “heads” or “tails”, because flipping a coin more closely emulates the outcome generation process.

Building on representativeness, we propose that prediction algorithms are judged to be more replicative when they capture the “essential properties” of the process that generates the outcomes to be predicted. However, the considerations that people make when judging replicativeness are likely different than those they make when judging representativeness because comparisons between outcomes and processes (regarding representativeness) are different from comparisons between two processes (regarding replicativeness). For example, different processes for predicting sequences of coin flips may rely on different instruments (e.g., flipping a coin, using a random number generator, or relying on a base rate), have different agents who act on the instrument (e.g., a computer program or a specific individual), and have different settings (e.g., the time and place in which the agent operates), even if the predictions that they produce are identical. Thus, all of these properties may affect people’s judgments of the replicativeness of different prediction algorithms, even if the predictions that these algorithms produce are equally representative of the process that generates event outcomes.

In what follows, we explain why people may prefer replicative prediction algorithms, in which domains replicative prediction algorithms perform poorly, and interventions that can lead people to choose more productive options when replicative algorithms fail.

Reasons For Belief in Prediction by Replication

Performance Expectations

People might prefer replicative algorithms because they expect replicative prediction algorithms to perform well. One reason behind this expectation is related to causal reasoning. When people observe an outcome, they often search for causes that might have led to the

outcome (Hastie 2015), and this tendency to engage in causal thinking is fast and spontaneous (Kahneman and Frederick 2002; Weiner 1985). When searching for causes, people often believe that antecedents of an event may have caused it (Hastie 2015; Lagnado and Sloman 2006). For example, a patient with a stomachache may naturally look to the most recent meal they ate as the potential cause of their illness.

Further, people often fail to consider randomness as a potential cause of the events that they witness (Hastie 2015). This tendency to neglect randomness in causal attributions means that people may naturally treat most prediction domains as though they are deterministic (see Einhorn 1986; Fox and Ülkümen 2011; Kahneman et al. 2021). Deterministic domains are those in which outcomes are unaffected by randomness, and the same process always leads to the same outcome. As evidence of this belief, people often underestimate the amount of variability in outcomes with the same inputs (see Clancy et al. 1981; Kahneman et al. 2021).

People's neglect of randomness as a cause may justifiably lead them to be optimistic about the performance of replicative prediction algorithms. In deterministic domains, replicative prediction algorithms do perform well because perfectly replicating all the antecedents of an event will always lead that event to produce the same outcome. For example, cooking a dish with the same exact conditions and procedures will always produce the same result, following the same route will always lead to the same destination, etc. Thus, people's experiences in deterministic domains might reinforce the belief that replicative algorithms perform well. To the extent that people have experienced success with replicative algorithms in the past (e.g., using recipes to cook food, learning the steps to solve a type of math problem), they may learn the heuristic that "following the same process leads to the same outcome".

However, replicative prediction algorithms are often an unwise choice in non-deterministic domains that are subject to random chance. The random component of these domains can play out differently every time that an outcome is generated, but a replicative method would attempt to duplicate the randomness inherent to the prediction domain, which typically reduces prediction performance. Probability matching is one classic example of a replicative prediction process that leads to suboptimal performance (Estes and Straughan 1954; McCracken et al. 1962; Shanks et al. 2002; West and Stanovich 2003). Probability matching describes people's tendency to predict outcomes with the same frequency that they occur. For example, a probability matcher predicting an event that results in "A" 70% of the time and "B" 30% of the time would essentially replicate the process that generates outcomes by picking "A" 70% of the time. This strategy leads to 58% accuracy, while relying on base rates and always predicting "A" leads to 70% accuracy. Similarly, the randomness inherent in driving, hiring, investing, and people's preferences means that following the same decision process can lead to different arrival times, hiring outcomes, investment returns, and success in selecting a product that a consumer will enjoy. However, people may generally apply the heuristic that the same process leads to the same outcome to deterministic and non-deterministic domains alike and favor replicative algorithms regardless of the prediction domain.

Liking of Replicative Algorithms Beyond Performance Expectations

Another reason that people may naturally be partial to replication as a means of prediction is that imitation, the act of replicating the behaviors of others, is key to human development (Bandura et al. 1961). Developmental psychology literature has found that individuals often use imitation as a tool when they are trying to solve a problem or make a decision. Specifically, people have the tendency to imitate all of the actions that an agent took to

achieve a goal when they are trying to achieve that same goal, regardless of how necessary the actions seem (Hoehl et al. 2019). For example, after observing another person opening a transparent puzzle box with necessary actions (e.g., inserting the key to the actual lock) and unnecessary actions (e.g., inserting the key to a visibly blocked hole), both children (Horner and Whiten 2005) and adults (Whiten et al. 2016) tend to mimic both the necessary and unnecessary actions when tasked with opening the same box. Thus, this literature suggests that people may see replication as a viable decision-making tool.

H1: The more replicative prediction algorithms are, the more people prefer to use them.

Boundaries of the Preference for Replicative Algorithms

Although we believe that replicativeness is generally desirable in prediction algorithms for the reasons described above, we also believe that there may be limits to the effects of replicativeness on people's preferences. As we discussed previously, one reason why people may like replicative prediction algorithms is that they might infer that more replicative algorithms will perform better. If this is the case, then people may rely less on replicativeness to choose among algorithms when they have performance information. That is, when people receive performance feedback about a prediction algorithm, they may learn to rely on that explicit performance feedback instead of less direct performance cues like replicativeness (Balzer et al. 1989; Shanks et al. 2002).

However, as proposed above, people may also have an innate preference for replicative prediction algorithms in addition to using replicativeness as a performance cue. As a result, we propose that making a prediction algorithm more replicative may generally lead people to like it more while holding its performance constant. For example, increasing people's perceptions of

the replicativeness of a navigation algorithm may generally increase their likelihood of using it even when they don't observe an improvement in the performance.

H2: People will shift to choose better performing algorithms when there is explicit performance feedback, but they will still prefer more replicative algorithms holding performance constant.

Because reminding people to pay attention to performance may not be enough to discourage the use of replicative algorithms in random domains, we propose an alternative intervention. We conjecture that if people like algorithms that replicate processes that produces events once, they may also like algorithms that replicate processes many times as long as each replication goes through a replicative process. We refer to algorithms that replicate a process many times and make predictions based on the aggregate results as "simulations" or "simulative algorithms".

In addition to being perceived as replicative, simulations also perform more accurately than algorithms that only replicate a process once. Because simulations repeat the event multiple times, they rely on a larger number of observations to generate predictions and are less prone to overfit any particular outcome. Furthermore, these algorithms calculate the best result over many trials so that they are less sensitive to randomness that may influence the outcome of any single event. Thus, simulative algorithms have the potential to benefit from people's desire for replicativeness while offering much better performance than algorithms that only replicate the generation process one time (or few times).

H3: People prefer replicative algorithms that rely on many replications of the predicted event.

Study Overview

In what follows, we test a number of empirical predictions derived from these hypotheses. In Study 1, we manipulate the replicativeness of prediction algorithms and find that people like a prediction algorithm more when they perceive it to be more replicative. In Study 2, we replicate our findings in multiple real-world domains. In Study 3, we find that performance feedback can drive people to choose algorithms based on performance, but people still prefer more replicative prediction algorithms holding performance constant. In Study 4, we propose simulation as an intervention that both performs well and satisfies people's desire for replicativeness, and test people's preference for a simulative algorithm[9].

For all studies, we set the sample size, exclusion criteria, and analysis plan before data collection. We report all exclusions (if any), all manipulations, and all measures in the manuscript, and additional information in the Web Appendix (accessible through OSF). In addition, we post preregistrations, study materials, data, and code for all studies in the online OSF materials (https://osf.io/r32wf/?view_only=4e6590694a404d1e878ab7aed6a6aa7b).

Study 1: Prediction of Die Roll Outcomes

In Studies 1a and 1b, we aim to test whether people prefer replicative prediction algorithms. We offer participants an algorithm that is replicative of process that generates event outcomes, and an "optimal" algorithm that gives the best answer for maximizing participants' incentive but is not replicative. Thus, we interpret any preference that participants express for the replicative algorithm in these studies as conservative because participants need to pay a cost to choose the more replicative algorithm.

[9] We used the term "algorithm" when describing participants' options in Studies 1a, 1b, & 4, and the term "method" when describing participants' options in Studies 2a-2c & 3. We also use the term algorithm in Study S1, S2 and S3. We found that participants' preference for replicative algorithms was consistent regardless of the terminology used. For consistency, we will describe all algorithms as "prediction algorithms" in the paper.

These studies use a simple prediction task so that the optimal prediction strategy and the process that generates outcomes in the task are clear. We apply what we learn in these simplistic studies to more externally valid prediction tasks in Studies 2a - 2c. The task used in Studies 1a and 1b asked participants to predict the number that a seven-sided die (with faces 1, 2, 3, 4, 5, 6, 7) will land on. We chose a seven-sided die roll for these studies because one number (4) minimizes average prediction error and maximizes participants' incentive, and rolling a die has a clear process.

We pre-registered that we would recruit 600 and 400 participants from MTurk for Study 1a and 1b respectively. Seven hundred and twenty-nine and 469 MTurk workers responded to the surveys, 604 and 405 of whom finished them and passed the attention checks. Participants' average ages were 37 and 36 (range: 18-80 and 19-70), and 47.2% and 44.7% were females.

Study 1a: Three Levels of Replicativeness

Procedures. Participants in Study 1a predicted the outcome of a seven-sided die roll (with faces 1, 2, 3, 4, 5, 6, 7). They faced a linear incentive scheme: they earned $0.21 if their chosen algorithm made a perfect prediction, and this bonus decreased by $0.03 for every unit of error in their chosen algorithm's prediction. Participants were randomly assigned to choose between two of three algorithms to make their prediction: a Replicative Algorithm, a Partially Replicative Algorithm, and an Optimal Algorithm.

The Optimal Algorithm always predicts that the outcome of the die roll will be 4, which is the number that minimizes absolute average error, and thus, maximizes participants' bonuses in expectation. The Replicative Algorithm generates a random number between 1 and 7 and submits that number as its prediction. Although the Replicative Algorithm uses a process that is more replicative of rolling a die, it also generates less accurate forecasts on average (e.g., it may

predict 7 when the outcome is 1). Thus, choosing the Replicative Algorithm is costly. The Partially Replicative Algorithm first flips a coin. If the coin lands on heads, it predicts that the outcome of the die roll will be 4. If it lands on tails, it generates a random number between 1 and 7 as its prediction. The Partially Replicative Algorithm performs better than the Replicative Algorithm because the former predicts 4 more often. It is also more replicative than the Optimal Algorithm but less replicative than the Replicative Algorithm because it matches the process of a die-roll half of the time. We included the Partially Replicative Algorithm to rule out the possibility that people simply dislike algorithms that always make the same prediction (e.g., the optimal algorithm that always predicts "4"). Overall, we expected participants to prefer the more replicative algorithm they were offered (replicativeness ordering: Replicative > Partially Replicative > Optimal) even though the more replicative algorithm always performed worse in expectation (earnings: Optimal > Partially Replicative > Replicative).

Participants were randomly assigned to make a binary choice between two of these algorithms (Optimal & Replicative, Replicative & Partially Replicative, or Optimal & Partially Replicative). After participants made their choice, the survey played out the die roll by drawing a random number between 1 to 7 and calculated their payment. Finally, all participants completed two exploratory questions that asked 1) how much they would earn when using each algorithm and 2) to explain the reason for their choice with an open text box. They then reported their age and sex to complete the study.

Results. We found support for the notion that participants prefer replicative prediction algorithms (H1). Overall, 57.8% (348/602) of participants chose the more replicative of the two algorithms they were offered, and the percentage of participants who selected the more replicative algorithm did not significantly differ across the three conditions (χ^2(2, N=602) =

0.41, p = 0.81). Expressing this preference for replicative prediction algorithms was costly for participants because the more replicative algorithm that participants were offered always earned less money in expectation. Further, the less replicative algorithm always stochastically dominated the more replicative option, so risk attitude alone cannot justify choice of the more replicative algorithm.

Although we did not preregister this analysis, we did find that participants chose the more replicative algorithm significantly more often than random chance (57.8% vs 50%), ($\chi^2(1)$ = 14.7, p < .001, ϕ_{cramer} = 0.16). This speaks against the alternative hypothesis that participants were simply choosing randomly. However, we do not believe that 50% is the right benchmark, as participants who wanted to maximize their incentive should have picked the more replicative algorithm 0% of the time (suggesting that 0% is a meaningful benchmark as well). However, 0% is also a problematic benchmark because any random answering will lead to greater than 0% choice of the more replicative algorithm. As a result, we preregistered that we were interested in whether "a substantial percent" of participants select the replicative algorithm. We do not have a formal definition of what constitutes "a substantial percent", so in this study, we report the percent of participants who chose the more replicative algorithm and leave this judgment to the reader (is it our subjective judgment that all reported percentages are "substantial"). For reference, we include all Chi-square tests against 50% in the footnotes.

Turning to the individual conditions, participants who chose between the Replicative and Optimal Algorithms chose the Replicative Algorithm most often (58%, 116/200) even though the Optimal Algorithm offered a larger bonus on average. Similarly, 59.3% (118/199) of participants chose the Partially Replicative Algorithm when the other option was the Optimal Algorithm.

These results are consistent with the notion that participants were more likely to pick prediction algorithms that more closely replicated the die roll, and thus, were more replicative.

One alternative hypothesis for not choosing the Optimal Algorithm is that participants simply don't like prediction algorithms that always select the same answer (i.e., "4"). To alleviate this potential concern, we included the Partially Replicative Algorithm that had a positive probability of selecting each possible outcome (1-7). In contrast to this potential concern, participants who chose between the Replicative Algorithm and the Partially Replicative Algorithm chose the Replicative Algorithm more often (56.2%[10], 114/203), even though the Partially Replicative Algorithm performs better on average. Overall, all of the results of Study 1 support the notion that people may prefer replicative prediction algorithms even when it is costly.

One open question after Study 1a is whether participants who chose the more replicative algorithm were capable of understanding that their chosen prediction algorithm generates lower bonuses on average than the alternative. We ran Study S1 (see Web Appendix D) to answer this question. In Study S1, we included a comprehension question that asked which bet (numbers 1-7) would result in the highest payoff on average. Overall, 62.7% of participants (252/402) answered the comprehension question correctly. For the condition in that study that most closely replicated Study 1a (in which participants answered the comprehension check after making their choice), we found that 39.8% (51/128) of participants who were able to answer this comprehension question correctly chose the more replicative algorithm anyway. This result

[10] 58% chose Replicative when choosing between Replicative and Optimal (116/200, $\chi^2(1) = 5.12$, p = .02, ϕcramer = 0.16). 59.3% chose Partially Replicative when choosing between Partially Replicative and Optimal (118/199, $\chi^2(1) = 6.88$, p = .008, ϕcramer = 0.19). 56.2% chose Replicative when choosing between Replicative and Partially Replicative (114/203, $\chi^2(1) = 3.07$, p = .08, ϕcramer = 0.12)

suggests that participants' preference for more replicative algorithms was not simply due to an inability to understand the costs of choosing the more replicative algorithm.[11]

Study 1b: Rolling a Die Vs Drawing a Marble

In Study 1b, participants judged multiple prediction algorithms that exhibit many different levels of replicativeness. In line with H1, we anticipate that people will prefer prediction algorithms that they judge to be more replicative. We identify two dimensions of a die-roll to replicate: the selection of outcomes from a uniform distribution and the action of rolling a die. We designed prediction algorithms that replicated neither, one, or both of these dimensions, measured participants' perceptions of replicativeness for each algorithm, and measured participants' preference for each algorithm.

Procedures. Participants in Study 1b once again predicted the outcome of a 7-sided die roll. They viewed five prediction algorithms, which were designed using a 2 (Action: match, no match) x 2 (Probability: match, no match) +1 (optimal) within-subjects design. "Matching the action" signifies that the prediction algorithm adopts the action of rolling a die, while "matching the probability" signifies that the prediction algorithm has the same probability distribution of outcomes as the die roll (i.e., a uniform distribution from 1 to 7).

The details of the five algorithms are as follows: the Optimal Algorithm always predicts that the outcome of the die roll will be 4. The Action-match Algorithm rolls an 11-sided die with the values 1, 2, 3, 3, 4, 4, 4, 5, 5, 6, 7 on each side, so that it matches the action of rolling a die but lacks the same probability distribution as the event being predicted. The Probability-match Algorithm draws one marble from an opaque jar with seven marbles of different colors

[11] Our exploratory measure of expected return found that people thought their chosen algorithm would perform better ($\beta = 0.036$, $t(599) = 11.18$, $p < .001$), and no algorithm differences after controlling for choice ($\beta = -0.00098$, $t(599) = -0.494$, $p = 0.62$) which means the expectation of performance is already captured in their choice.

corresponding to the numbers 1 through 7, so that the probability distribution matches the event being predicted (1-7 uniform), while the action (drawing a marble) does not match. The Perfect-match Algorithm rolls a 7-sided die so that both the action and the probability distribution match the event being predicted. The No-match Algorithm draws a marble from an 11-marble jar with marbles corresponding to the values 1, 2, 3, 3, 4, 4, 4, 5, 5, 6, 7, so that neither the action nor the probability distribution matches the event being predicted. The exact wording of the stimuli is displayed in Web appendix A. According to our theorization, participants will like the Perfect-match algorithm the best because it best replicates the die-roll on both the probability distribution and the action. We also propose that participants will prefer the Action-match Algorithm and Probability-match Algorithm to the Optimal Algorithm and No-match Algorithm because one dimension of the event is more closely replicated.

Participants in Study 1b were asked to imagine they would get a bonus according to the linear incentive scheme from Study 1a, earning $0.21 if their chosen algorithm made a perfect prediction $0.03 less for every unit of error in their chosen algorithm's prediction. As a result, "4" (i.e., the Optimal Algorithm) was once again the option that minimized average error and maximized participants' bonus. Also, the algorithms that did *not* "match probability" (i.e., didn't have a uniform distribution) performed better than those that did. All participants rated their likelihood of using each algorithm to predict a die roll on a 5-point Likert scale (1= *Extremely Unlikely* and 5 = *Extremely Likely*). Participants made these ratings in randomized order and on a separate page for each algorithm.

After participants rated their likelihood of using each algorithm, they rated two measures of perceived replicativeness: "To what extent do you agree that the algorithm goes through the same process as rolling a die?" and "To what extent do you agree that the algorithm and the activity

that you are predicting generate the same outcomes?" Participants rated both of these questions for all five algorithms on 5-point Likert scales. Both of the questions were presented on the same page for each algorithm, and the pages were presented in random order. They then reported their age and sex to complete the study.

Results. As pre-registered, we took the average of the two replicativeness ratings (α = 0.77, *95% CI* = [0.75, 0.79]) as a composite measure of replicativeness. As a manipulation check, we first examined the perceived replicativeness rating of each algorithm. For each algorithm excluding the Optimal Algorithm, we contrast coded whether it matches the action of rolling a die (Matching Action = 1, Not Matching Action = -1) and whether it has a uniform probability distribution matching the die roll (Matching Probability = 1, Not Matching Probability = -1). Then, we regressed the average replicativeness rating for each algorithm onto the contrast-coded indicators of action and probability matching, and the interaction between the two. We clustered standard errors by participant IDs to account for the repeated ratings. Participants' perceived replicativeness of the prediction algorithms increased when the algorithm matched the probability distribution of the prediction event ($\beta = 0.46$, $t(401) = 15.13$, $p < .001$) and when the algorithm matched the action of rolling a die ($\beta = .14$, $t(401) = 7.04$, $p < .001$). There was also no significant difference in the replicativeness rating between the Optimal Algorithm and the No-match Algorithm (2.40 vs 2.76, $t(404) = 1.09$, $p = .28$), suggesting that algorithms that do not match either dimension are perceived to be similarly non-replicative. We also found a positive effect on the interaction between action and probability ($\beta = 0.03$, $t(401) = 2.09$, $p = .036$). This interaction suggests that people's perceptions of replicativeness may follow a multiplicative rule: matching both the uniform distribution of a die roll and the action of rolling a die has an additional positive effect beyond the sum of the independent effects.

We found similar effects of our manipulation on participants' likelihood of using the algorithms. We regressed participants' likelihood-to-use ratings onto contrast coded predictors of whether each algorithm matched the uniform distribution and the action of rolling a die, excluding the Optimal Algorithm, and clustered standard errors by participant ID. Participants rated an algorithm higher when it matched the probability distribution of the prediction event ($\beta = 0.26$, $t(401) = 8.51$, $p < .001$) and when it matched the action that generated event outcomes ($\beta = 0.14$, $t(401) = 6.11$, $p < .001$). In addition, there was a positive interaction between probability and action matching ($\beta = 0.09$, $t(401) = 4.56$ $p < .001$). Further, participants rated the Perfect-match Algorithm higher than all other algorithms (t'$s(404) \geq 7.16$, p'$s < .001$; figure 1). Overall, these results support H1, suggesting that people are more likely to use prediction algorithms that they perceive to be more replicative.

To further examine the relationship between replicativeness and participants' likelihood of using a prediction algorithm, we investigated if participants' perceptions of replicativeness mediate their rated likelihood of using a prediction algorithm. We performed two OLS regressions predicting the algorithm ratings with five condition dummies, standard errors clustered by a participant ID, and no intercept. One included the average replicativeness ratings as a mediator, and the other did not. We tested the joint hypothesis that the coefficients of the five dummies are equal, and found the ratings were significantly different across conditions ($F(4, 404) = 29.85$, $p < .001$). After including the replicativeness ratings, we found that the F-value dropped 21.01 points ($F(4, 404) = 8.84$, $p < 0.001$), and a bootstrapped 95% confidence interval of the drop excludes zero (11.07, 36.14). This significant drop in F-value indicates participants' replicativeness ratings mediate the differences in preference ratings between prediction algorithms. We interpret this as

correlational evidence that that the more replicative a prediction algorithm is of the predicted event's process, the more people will want to use it.

Discussion

Studies 1a, 1b, and S1 found preliminary evidence that many people prefer replicative prediction algorithms even when choosing those algorithms is costly. The studies provide evidence against the alternative explanations that people simply dislike prediction algorithms that always make the same prediction, and that participants were simply not able to understand the costs of choosing the more replicative algorithm. Further, these results suggest that people's perceptions of replicativeness are at least somewhat continuous, and that people prefer prediction algorithms that they perceive to be more replicative.

The results of Study 1b also demonstrate the importance of understanding the replicativeness of prediction algorithms, and not just the representativeness of the predictions that they produce. Both the Probability-match and Perfect-match algorithms produced predictions that were representative of the seven-sided coin flip being predicted – draws from a uniform 1-7 distribution. Consistent with a preference for representative predictions, participants preferred both of these algorithms to otherwise identical algorithms that made different predictions. However, participants also preferred the Perfect-match algorithm to the Probability-match algorithm even though both make the same predictions, and thus, produce equally representative outcomes. We found evidence that this preference for the Perfect-match algorithm was due to its replicativeness; participants preferred algorithms that replicated the process of rolling a die (instead of following non-replicative algorithms like drawing marbles or always making the same prediction). This suggests that people's preference for replicative algorithms is

not simply based on the representativeness predictions that they produce, but also the replicativeness of the processes that they use to produce those predictions.

Study 1 found evidence for people's preference for replicative prediction algorithms. We found this preference in a relatively artificial domain with a clear process and optimal answer. In Studies 2a-2c, we test whether people prefer replicative prediction algorithms in more externally valid real-world domains.

Study 2: Prediction in Real-world Domains

Study 2 tests people's preference for replicative prediction algorithms in real-world domains. Specifically, we examine people's preference for replicative prediction algorithms in sports betting (Study 2a), movie recommendations (Study 2b), and clothing subscription services (Study 2c). We pre-registered that we would recruit 400, 200, and 800 participants from MTurk for Study 2a, 2b, and 2c respectively. Four hundred and thirty-one, 225, and 850 MTurk workers responded to the surveys, 403, 200, and 802 of whom finished them and passed the attention checks. The average ages in the final samples were 38.0, 38.0, and 41.3 (ranges: 20-79, 19-71, and 19-79), and 47.1%, 48%, and 45.8% were females.

Study 2a: NBA Matchups

Procedures. In Study 2a, participants learned that they would choose between two real prediction algorithms to predict the winners of nine NBA matchups on the day after the experiment, and they received a 10-cent bonus for each matchup that their chosen algorithm predicted correctly. Both algorithms were hosted on Nate Silver's fivethirtyeight.com at the time of the experiment, and are still hosted there at the time of this writing (Silver et al. 2021). The two algorithms were labelled Method E (team-based Elo metric on 538) and Method R (individual-based RAPTOR metric on 538). Participants read that Method E "is based entirely

on team metrics (i.e., it doesn't account for individual player metrics). A given team's rating is generated by aggregating past team data like margins of victory, home court advantage, and the quality of opponents", and Method R "is based entirely on NBA player projections (i.e., it doesn't account for wins, losses, and other team metrics). These player projections estimate each player's future performance based on the trajectory of similar NBA players." The individual-based algorithm (Method R) is used as the default on fivethirtyeight.com because of its performance advantage (Silver et al. 2021). However, we believed that participants would perceive the team-based algorithm (Method E) to be more replicative of the process that determines basketball game outcomes because basketball is commonly represented as a team-sport, and we verified this assumption in a pretest (See Web Appendix B).

Participants were assigned to one of two conditions. Half of the participants were assigned to the *team-framing* condition, in which they predicted the winners of the games. The other half of participants were assigned to the *individual-framing* condition, in which participants instead predicted which set of individual players will score more points per minute on average. The prediction problem was identical between the two conditions, as the set of players who score more points per minute on average win the game. However, we expected that manipulating the framing would change participants' representation of basketball games. Specifically, we expected that *individual-framing* would emphasize the role of individual performance in basketball game outcomes and consequently increase the perceived replicativeness of the individual-based algorithm. Importantly, the description of the algorithms was the same in both conditions, and we only changed the framing of the event being predicted. This allows us to manipulate the extent to which these algorithms are perceived to replicate the event holding all else constant and get at the causal relationship between replicativeness and preference.

After making their choice, participants rated their agreement with two replicativeness measures on a 5-point scale (1 – Strongly disagree; 5 – Strongly agree), with the exact framing altered depending on their condition. The measures in the *team-framing condition [individual-framing condition]* are: 1) This method shares similarities with the process that determines *the outcome of NBA basketball games* [*players' average rate of scoring*] and 2) The way that *a team wins a matchup* [*a set of players scores in the NBA*] is well represented by this method. Participants rated both of the questions on the same page for each algorithm, and the order of the algorithms was randomized. Participants reported their age and sex to complete the study.

Results. As pre-registered, we took the average of the two replicativeness ratings (α = .71, *95% CI* = [.68, .74]) as a composite measure of replicativeness. Consistent with the pretest, participants in the *team-framing* condition rated the team-based algorithm as more replicative than the individual-based algorithm (M_{team} = 4.00, $M_{individual}$ = 3.11, $t(200) = 9.35$, $p < .001$) and chose it more often than the individual-based algorithm (72.1%, compared to 50%: 145/201, $\chi^2(1) = 39.4$, $p < .001$, $\phi_{cramer} = 0.44$). However, difference in choice could be due to other differences between the algorithms beyond replicativeness. Thus, it is essential to examine the manipulation of replicativeness as this analysis holds other differences between algorithms constant. Participants in *the individual-framing* condition rated the individual-based algorithm as more replicative than participants in the *team-framing* condition (M_{team} = 3.04, $M_{individual}$ = 3.87, $t(201) = -8.27$, $p < .001$), and the proportion choosing the individual-based algorithm rose from 27.9% in the *team-framing* condition to 41.6% in the *individual-framing* condition. (χ^2 (1) = 7.78, $p = .005$, $\phi_{cramer} = 0.14$). These results suggest that we successfully manipulated the individual-based algorithm to seem more replicative, and that this increase in replicativeness led participants to choose the individual-based algorithm more often. Overall, excluding participants

who rated the two algorithms as being equally replicative, 73.9% (226/306, χ^2 (1) = 69.66, p < .001, ϕ_{cramer} = 0.48) of participants chose the algorithm that they rated as more replicative, consistent with the notion that people tend to prefer replicative prediction algorithms.

Study 2b: Movie Ratings

Procedures. Participants in Study 2b chose between two algorithms to predict a person's (the predictee's) rating of a movie. The predictee and the movie were randomly selected from a pre-collected movie rating database. The database consists of 946 MTurk workers' ratings of 50 famous movies and 14 movie genres.

Participants could choose between two prediction algorithms. If the algorithm they chose performed better than or equally to the alternative (i.e., was closer to or equally distant from the predictee's actual rating), they would get a $0.25 bonus. The Replicative Algorithm finds the person (Person B) in the database who rated the 14 genres most similarly to the predictee and has watched the target movie. It then submits Person B's rating of the target movie as its prediction of the predictee's rating. We propose that this algorithm is more replicative because a person who has the most similar taste may be most likely to go through the same judgment process as the predictee when rating a movie. The less replicative Average Algorithm submits the average rating of the target movie among all people in the database as its prediction of the predictee's rating. We propose that this algorithm is less replicative because it takes all the people in the database into account instead of only using people with similar preferences. Although the Average Algorithm is less replicative, it actually performs better (i.e., it provides a rating closer to the predictee's actual rating 1.5 times more often in our database) because it does not overfit any one participant's responses. After participants made their choice, they answered an

exploratory question asking the reason for their choice with an open text box and reported their age and sex to complete the study.

Results. Even though the Average Algorithm is more accurate, 71% (142/200, $\chi^2(1) = 35.28$, $p < .001$, $\phi_{cramer} = 0.42$) of participants chose the Replicative Algorithm for their prediction, consistent with the notion that people prefer replicative prediction algorithms (H1). We further replicate these findings in another consumer domain: evaluations of navigation algorithms (Study S2 in Web Appendix D).

Study 2c: Clothing Subscriptions

Procedures. In Study 2c, participants rated different descriptions of clothing recommendation systems on a number of dimensions including willingness-to-use, replicativeness, understanding, perceived cost, and uniqueness. We investigated whether participants' perceived replicativeness of the recommendation systems predicted their willingness to use each algorithm.

Participants viewed descriptions of a clothing subscription service's recommendation system and rated them on a number of dimensions. Each participant viewed four different descriptions, randomly selected out of a pool of eight, adopted and paraphrased from a clothing subscription company's (StitchFix) actual description of their algorithm (Colson et al. 2021) (Table 1). Our goal in designing these stimuli was to see if subtle variations in a prediction algorithm's description can change people's perceived replicativeness, and their willingness to try that algorithm as a result. For the dependent variable, participants rated "To what extent are you willing to try the subscription service?" on a 7-point likert scale (1 = Don't want to try at all; 7 = Can't wait to try) for each of the recommendation systems. After rating the willingness-to-use for all four prediction algorithms, participants further rated each algorithm on several other

dimensions using 5-point likert scales. The dimensions included how replicative the algorithm was of the process that they use to pick their own clothes, how expensive they perceived the service was, how unique the algorithm was, and how well they understood the algorithm (see Table 2 for the exact wording of all questions). Participants rated all these dimensions for one algorithm before proceeding to the next algorithm, and the order of these dimensions and the order of algorithms were both randomized. We included dimensions other than replicativeness so that our hypothesis was not obvious to participants, and so that we could test whether participants prefer replicative algorithms after controlling for these other dimensions. Participants reported their age and sex to complete the study.

Results. As pre-registered, we took the average of the two replicativeness ratings replicativeness (α = 0.91, *95% CI* = [0.90, 0.91]) as a composite measure of replicativeness. There was a positive correlation between the replicativeness ratings and willingness-to-use ratings (Pearson's r = 0.22, $t(3206) = 12.71$, $p < .001$; figure 2), suggesting that participants preferred algorithms that they perceived to be more replicative. Next, we test the effect of replicativeness when controlling for all other dimensions. We regressed the willingness-to-use ratings of the eight services onto all rated dimensions with standard errors clustered by participant IDs. This analysis returns a significant effect of replicativeness (β = 0.32, $t(797)$ = 6.73 , $p < .001$), which suggests that the construct of replicativeness is distinct from uniqueness, perceived price, and understanding, and that participants valued replicativeness even after accounting for these dimensions (see Web Appendix C for full regression outputs). This effect remains significant after adding fixed effects for all eight algorithms (β = 0.31, $t(790)$ = 6.35, p

$< .001$)[12]. These results are consistent with the notion that the more replicative participants thought a prediction algorithm was, the more willing they were to use it.

Next, we investigated if participants' perceptions of replicativeness mediated their reported likelihood of using the recommendation system. We performed two OLS regressions predicting the willingness-to-use ratings with eight description dummies, standard errors clustered by participant IDs, and no intercept. One included the replicativeness as a mediator, and the other did not. We tested the joint hypothesis that the coefficients of the eight dummies are equal, and found that the ratings were significantly different across conditions ($F(7, 801) = 6.70$, $p < .001$). After including the replicativeness ratings, we found that the F-value dropped 3.70 points ($F(7, 801) = 3.00$, $p = .004$), and a bootstrapped 95% confidence interval of this drop excludes zero (1.96, 5.97). This significant drop in *F*-value (i.e., mediation) provides correlational evidence that that the more replicative a prediction algorithm is of the predicted event, the more people prefer it, which is consistent with H1.

Discussion

Consistent with the results of Studies 1a and 1b, the results of Studies 2a - 2c support the notion that people prefer prediction algorithms that are more replicative (H1). Studies 2a - 2c & S2 specifically provide evidence that people prefer more replicative prediction algorithms in more real-world domains than Studies 1a - 1c. We found evidence that people's preference for prediction by replication applies to prediction algorithms with both objective outcomes (e.g., outcomes of basketball games; Study 2a) and subjective evaluations (e.g., consumers'

[12] The coefficient of replicativeness is similar and significant using random effects (see Web Appendix C).

preferences for clothes; Study 2c), which suggests that this preference could be somewhat robust across contexts.

The results so far provide many implications for practitioners designing and promoting prediction algorithms. First, we found that subtle changes to descriptions of prediction algorithms (recommendation systems in Study 2c) and prediction events (basketball games in Study 2a) can lead to changes in the perception of replicativeness of prediction algorithms and consequently changes in consumers' willingness to use the algorithms. These results suggest that practitioners may often be able to simply reframe descriptions of current prediction algorithms and events to improve consumers' evaluations of those algorithms. As changing these descriptions is relatively costless (compared to changing the actual prediction algorithm), it may be hugely beneficial for practitioners to learn how to frame their prediction algorithms to be as replicative as possible.

The results of Studies 2b & S2 suggest that it may be persuasive to frame the cases that predictions are based on to be as similar as possible to the case being predicted. For example, when predicting which songs a customer will enjoy, it may be persuasive to describe the data that these predictions are based on as coming from customers who have expressed nearly identical preferences for music in the past. Based on previous work on the law of small numbers, consumers are unlikely to recognize the downside of basing predictions on only the most similar cases – small samples sizes (Tversky and Kahneman 1971). Consistent with this thinking, the majority of participants in Study 2b favored a prediction algorithm based on the one most similar case. Although basing prediction algorithms on small samples sizes is likely to lead to inaccurate predictions, practitioners can frame predictions from algorithms that account for a multitude of factors (e.g., a regression with many dependent variables) as using data from the most similar

cases or providing predictions that are particular to a customer's unique traits to leverage people's preference for replicativeness.

Study 3: Multiple Predictions and Choices

In Study 3, we test a potential boundary of people's preference for replicative prediction algorithms. We propose in H2 that people who receive performance feedback may shift to choosing better performing algorithms, and place relatively less weight on replicativeness. Study 3 tests this prediction using the die-roll paradigm from Study 1 by asking participants to make multiple decisions and providing them performance feedback. In addition to providing performance feedback, we manipulate both the replicativeness and the performance of the same target prediction algorithm. This allows us to isolate the effects of both replicativeness and performance and test whether people simultaneously prefer better performing algorithms, and more replicative algorithms holding performance constant (H2).

Participants and Procedures. We pre-registered that we would recruit 800 participants. Eight hundred and twenty-eight MTurk workers responded to the survey, 802 of whom finished it and passed the attention check. The average age in the final sample was 41.0 (range: 18-89), and 45.4% were females.

Participants in this study made 20 choices between two algorithms to predict 20 outcomes of a 7-sided die-roll (similar to Study 1). The study used a 2 (replicativeness: target-more-replicative; target-less-replicative) by 2 (accuracy: target-better, target-equivalent) between-subject design. The target algorithm in all conditions "rolls a die with the values 1, 2, 3, 4, 4, 4, 5, 6, 7 and submits the die roll's outcome as its prediction". To manipulate replicativeness, we paired the target algorithm with different alternative prediction algorithms (Table 3). For the target-less-replicative condition, participants chose between the target

algorithm and a more replicative algorithm that "rolls a die with the values 1, 2, 3, 4, 5, 6, 7" and submits the die roll's outcome as its prediction. As shown in Studies 1a & 1b, people perceive an algorithm that draws from the same distribution of outcomes as the prediction event to be more replicative and prefer it as a result. For the target-more-replicative condition, participants chose between the target algorithm and a less replicative algorithm that "predicts that the outcome of the die roll will be 6".

To manipulate accuracy, we randomized the bonus scheme assigned to participants. In the target-equivalent condition, participants got a bonus of 0.07 per trial only for a perfect prediction, and hence both algorithms perform equivalently because all algorithms make a perfect prediction $1/7^{th}$ of the time. Therefore, choosing the more replicative algorithm does not have a performance advantage over any of the alternative algorithms, and this condition allows us to examine participants' preferences when performance is held constant. In the target-better condition, participants got a bonus based on how close their prediction was to the outcome. For each prediction, they got a 0.07 bonus if they predicted perfectly, and the bonus decreased by one cent with each unit of error. Given this incentive, the number that yields the highest bonus on average is 4, and the target algorithm performs better (off by 2.16 on average) than either alternative algorithm (which are both off by 2.29 on average). This condition allows us to examine how performance interacts with replicativeness. All conditions are described in detail in table 3.

Participants first read about the prediction problem and the bonus scheme. They learned that they would receive the sum of all bonuses from each of their 20 choices (up to $1.40). Then, they viewed the two algorithms, which appeared on the screen in random order, and chose between them. After they made a choice, the interface picked a random number between 1 to 7

as the outcome of the die roll. On the next screen, the interface showed what the right answer was, what each algorithm guessed, and how much bonus money they earned from that trial (see Web Appendix E for detailed stimuli). Participants repeated this choice and received feedback 20 times, and reported their age and sex to complete the study. With this multiple-choice set-up, participants were able to explicitly observe the performance of each algorithm and gain evidence that replicative algorithms don't perform better in the target equivalent conditions. Therefore, we can examine whether the preference for replicative algorithms still holds after people have evidence that those algorithms do not perform better than alternatives. Similarly, the condition with a linear incentive and a more replicative alternative algorithm allows us to test how people react to performance information regarding a more replicative algorithm (the alternative) that performs worse than a less replicative algorithm (the target).

Results and Discussion. As pre-registered, we performed a logistic regression with the dependent variable as whether the target algorithm was chosen. The predictors included whether the target was more replicative (1 – more replicative; 0 – less replicative), whether the target was more accurate (1 – more accurate; 0 – equally accurate), choice number (0-19), and all the two-way interactions between the three predictors. We clustered standard errors by a participant ID to account for repeated choices (full regression in Table 4)[13]. We found a significant effect of accuracy ($\beta = .38$, $z = 2.69$, $p = .007$, $OR = 1.46$), a significant effect of replicativeness ($\beta = .38$, $z = 2.74$, $p = .006$, $OR = 1.46$), and no significant interaction between the two ($\beta = -.003$, $z = -0.018$, $p = .99$, $OR = 1.00$). These simple effects indicated that participants preferred replicative

[13] We did not include a three-way interaction in our pre-registration, but adding the three-way interaction into the model does not meaningfully change the results (and the three-way interaction is not significant). Similarly, including random effects instead of clustered standard errors does not meaningfully change the pattern of results. Further, the results are virtually unchanged when accuracy and replicativeness are contrast coded (see Web Appendix E).

prediction algorithms after accounting for accuracy. The lack of interaction indicated that even when participants observe that the more replicative algorithm is no more accurate than the alternative, they still prefer it to a less replicative algorithm (figure 3). There was also no significant interaction between choice number and replicativeness ($\beta = .002$, $z = 0.37$, $p = .71$, OR = 1.00), suggesting that the preference for replicativeness was relatively consistent throughout the 20 trials. There was a significant positive interaction between choice number and accuracy ($\beta = .016$, $z = 2.63$, $p = .009$, $OR = 1.02$), indicating that when the target was more accurate, participants were more likely to choose it the more choices they made (and the more explicit performance feedback they got). In other words, as people receive more explicit performance feedback, the importance of a prediction algorithm's performance likely increases relative to its replicativeness. However, people's preference for replicativeness appears to be relatively stable.

To contrast participants' use of accuracy and replicativeness in the beginning and in the end of the 20 trials, we preregistered two logistic regressions predicting whether participants chose the target algorithm (0 = not chosen, 1 = chosen) for participants' first five and last five trials. The predictors were the replicativeness and accuracy of the target algorithm dummy coded (as they were above), with standard errors clustered by participant ID. We found that both accuracy ($\beta = .40$, $z = 3.83$, $p < .001$) and replicativeness ($\beta = .45$, $z = 4.28$, $p < .001$) had a significant effect on participants' first five choices. These effects were of a similar size ($z = -0.31$, $p = 0.76$), suggesting that these two factors received a similar weight in participants' decisions. The coefficient of the accuracy term marginally increased to 0.66 when predicting the last five choices (compared to $\beta_{first5accuracy}$, $z = 1.64$, $p = .10$), while the coefficient of the replicativeness term ($\beta = .48$, $z = 4.16$, $p < .001$) remained a similar magnitude ($z = .16$, $p = .87$).

These results suggest that both replicativeness and accuracy are important from the onset, but that people may value accuracy more over time while the effect of replicativeness is relatively consistent across multiple trials. Therefore, people may learn to weigh performance more than replicativeness with feedback, but consistent with H2, people prefer replicative algorithms holding all else equal even after getting performance feedback.

Discussion

Study 3 provides insight into the effect of replicativeness as people learn about an algorithm's performance over time. Consistent with H2, people shift to choose the better-performing method when performance feedback is explicitly available, but still prefer prediction algorithms to be replicative holding performance constant. The effect of replicativeness was relatively stable over time even as participants received feedback about a prediction algorithm's performance. In contrast, the effect of performance grew (marginally) over time as participants gained performance feedback.

The results of Study 3 suggest that people still prefer relatively replicative prediction algorithms holding all else equal after getting performance feedback. In Study 4, we aimed to test whether we could leverage people's preference for replicative prediction algorithms to get them to use an alternative that performs better. Specifically, we test the hypothesis that people may like prediction algorithms that perform a replicative process many times, which we dub "simulative algorithms". People may find simulative algorithms desirable because they can rely on the same replicative processes as prediction algorithms we have tested so far, but simulative algorithms will offer better performance because they do not overfit any one particular random error. For example, in predicting the outcome of a die roll, a simulative algorithm would roll the

die many times and find the number that yields best prediction on average. We investigate people's preference for simulative algorithms in Study 4.

Study 4: Simulation as Intervention

We designed Study 4 to test whether people still prefer better-performing replicative algorithms that simulate outcomes many times. We pre-registered that we would recruit 400 participants. Four hundred fifty-four subjects responded to the survey, 402 of whom finished it and passed the attention check. The final sample had an average age of 37 (range: 19-98) and 54.2% females.

In this study, we tasked participants with the same die-roll prediction problem as Study 1 and 3. We designed two different simulation algorithms. Both replicate the event of rolling a die 10 times but differ in how they calculate the prediction. We propose that because both algorithms involve the replication of the event, people will perceive both as replicative and be open to using either.

Procedures. Participants rated how likely they were to use four different on a 5-point scale (1 = *Extremely Unlikely* and 5 = *Extremely Likely*). They were instructed to imagine the same linear incentive scheme as in Study 1 (i.e., receiving $0.21 for predicting correctly, $0.18 for being off by 1, etc.). The four algorithms included the Optimal Algorithm and the Replicative Algorithm used in Study 1. The study also included two simulation algorithms. The Total Algorithm rolls a seven-sided die with 1, 2, 3, 4, 5, 6, 7 on each side 10 times. For each of the 10 rolls, the algorithm calculates how much money choosing each number (1, 2, 3, 4, 5, 6, 7) would have earned. Then, the algorithm determines which number would have earned the most money across all of the rolls and submits that number as its prediction. The Average Algorithm rolls a seven-sided die with 1, 2, 3, 4, 5, 6, 7 on each side 10 times. The algorithm then calculates the

average outcome of those 10 rolls and submits the number that is closest to the average as its prediction. Because the two algorithms both involve rolling a die 10 times, we believe both will be perceived as being replicative of the event in question. We included two simulation algorithms to explore if people's preference for simulation is robust to the way it is described. After rating their likelihood of using each algorithm, participants rated the replicativeness of each algorithm using the same questions as Study 1b.

Results and Discussion. As hypothesized, participants expressed significant differences in their willingness to use the four algorithms ($F(3, 1604) = 37.22$, $p < .001$). Unpacking the differences, we replicated the finding from Study 1 that people were more willing to use the Replicative Algorithm ($M = 3.12$) than the Optimal Algorithm ($M = 2.78$, $t(401) = 3.55$, $p < .001$). Further, consistent with our hypothesis, participants were more willing to use the Total Algorithm ($M = 3.52$) and the Average Algorithm ($M = 3.52$) than the Optimal Algorithm ($M = 2.78$, $t(401) = 8.62$ and 8.19, p's $< .001$), suggesting a preference for Replicative Algorithms that replicate the event multiple times. Participants also rated the simulation algorithms higher than the Replicative Algorithm ($M = 3.12$, Cohen's $d = 0.34$ and 0.34, $t(401) = 5.78$ and 6.05, p's $< .001$), which further suggests that simulation algorithms could be successful as an intervention. We also replicate this preference in Study S3 when asking participants to rate different algorithms for predicting roulette outcomes (Study S3 in Web Appendix D).

We also replicated the finding that perceived replicativeness mediates participants' ratings of their likelihood of using the different prediction algorithms. We first checked that participants rated the replicativeness of the four algorithms differently ($F(3,1604) = 85.76$, $p < 0.001$; $M_{optimal} = 2.44$ ($SD = 1.25$), $M_{replicative} = 3.54$ ($SD = 1.02$), $M_{total} = 3.24$ ($SD = 1.06$), $M_{average} = 3.45$ ($SD = 1.00$)). Then, we performed two OLS regressions predicting the

algorithm ratings with four condition dummies, no intercept, and clustering standard errors by participant ID. One included the replicativeness ratings as a predictor, and the other did not. Testing the joint hypothesis that the four conditions were equal in the two regressions, we found that the average replicativeness ratings for the algorithms mediated the difference in the likelihood-of-using between algorithms ($F_{noMediator}$ (3, 1604) = 37.22, $F_{mediator}$(3, 1604) = 25.71, drop in *F-value* = 11.51, *95% CI* = [3.03, 22.49]). These findings provide correlational evidence that people prefer simulative algorithms that replicate the task multiple times because they perceive those algorithms to be replicative.

General Discussion

In seven studies, we provide evidence that people prefer prediction algorithms that replicate the event being predicted. In Studies 1a, 1b, & S1, we found that people prefer replicative algorithms, and that more replicative people perceive an algorithm to be, the more they like to use it. Studies 2a through 2c & S2 show how such a preference may influences people's sport betting decisions, and selections of recommendation algorithms and navigation algorithms. In Study 3, we found that, although people learn to choose higher performing algorithms after getting performance feedback, they still prefer more replicative algorithms holding performance constant. Therefore, we propose simulation (Study 4) as an intervention that achieves good performance and leverages people's desire for replication. As predicted, we found that people did embrace simulative prediction algorithms.

Practical Implications

These findings imply that managers and marketers may often be able to boost adoption of prediction algorithms by describing them in replicative terms. Many prediction algorithms are already somewhat replicative by design or have features that can be framed as being replicative.

For example, weather forecasts generate predictions by simulating how clouds and the atmosphere are going to move (Lynch 2008), and thus, can be accurately framed as replicating the process that will generate future weather events. Likewise, recommendation systems based on collaborative filtering (a commonly used methodology, Schafer et al. 2007) pick products for a target consumer by looking up the past behavior of the most similar consumers, who may have decision processes that are most similar to the target. As these algorithms already have somewhat replicative processes, highlighting this replicativeness may be a relatively costless way to persuade customers to adopt the company's products or services.

While the processes that generate outcomes in some prediction domains are relatively transparent (e.g., the rules and procedures for playing blackjack), others are not (e.g., the process that determines stock market prices). Thus, in these opaque domains, understanding how to make a prediction algorithm more replicative will often come down to gaining a better understanding of people's beliefs. For example, marketers who want to frame a music recommendation system as being replicative first need to understand consumers' beliefs about how their music preferences are formed. More concretely, if consumers believe that their music preferences are driven by features of songs, marketers at Pandora could potentially make their system seem more replicative by more explicitly highlighting that their algorithm can find and recommend songs with the user's preferred features (see Pandora Media 2021).

People may also prefer a prediction algorithm when it is framed as accounting for the greatest number of factors possible even when some factors are limited in predictive power. To the extent that people believe that a prediction algorithm that incorporates more factors may be better able to emulate the event being predicted, they should also believe that prediction algorithms that incorporate more factors are more replicative. Therefore, even if developers learn

that a navigation app does not make more accurate predictions when it accounts for the day of the week and the previous day's traffic, it may be beneficial to let these factors play a minor role in the prediction process to be able to advertise that the application accounts for these additional factors. While statisticians are taught to build the simplest model possible holding performance fixed, this may not be the most persuasive way to build models when adoption is an objective.

Finally, this work suggests when it may be most productive to leverage replicativeness. The results of Study 3 suggest that the absolute effect of making an algorithm more replicative is relatively constant over time, but that people may pay relatively more attention to performance as they gain more performance feedback. This implies that designing or describing algorithms as more replicative may be most effective when attracting people to a new algorithm because messaging about replicativeness may get the most weight relative to other considerations before experiencing the prediction algorithm.

Theoretical Implications

Our findings suggest additional potential mechanisms for some established phenomena. First, people's desire for replicative prediction algorithms could also provide insight into when and why people prefer human decision makers to algorithms. For example, people often prefer human decision makers when the task is subjective (Castelo et al. 2019), when making medical decisions (Longoni et al. 2019), and when decisions are morally relevant (Dietvorst and Bartels 2021). It is possible that this preference for human decision makers in these domains could relate to people's desire for prediction by replication. That is, people may want humans instead of machines to predict outcomes that are in some way generated by humans, like subjective preferences, human health outcomes, and moral judgments. Further, people may be more

comfortable using algorithms to predict outcomes that are not generated by humans (see Castelo et al. 2019) because human judges may not be inherently more replicative.

Our results could also relate to work suggesting that similar others are often more persuasive than dissimilar others. A large body of literature shows the use of figures similar to consumers in advertisements (McGuire 1978) and sales forces (Woodside and Davenport Jr. 1974) can be more persuasive (Petty and Cacioppo 1986; Burger et al. 2004) and more effective in increasing consumers' purchase intentions (Jiang et al. 2010). These findings have often been attributed to consumers' liking of similar figures, but our work provides an additional account that may also drive these results: consumers may believe that similar others have decision processes that are closer to their own or may be better able to emulate their preferences.

Our findings may also connect to many established phenomena in psychology. For example, Gambler's fallacy is the belief that, in random domains, runs of a particular outcome (e.g., many heads when tossing a coin) will be balanced out by occurrences of alternate outcomes (e.g., tails; Ayton and Fischer 2004; Tversky and Kahneman 1971). At first blush, this seems contradictory to a preference for replicative forecasting algorithms, as the preference for replicativeness implicitly assumes that two outcomes of identical processes will match (instead of alternating as suggested by gambler's fallacy). However, in studies of gambler's fallacy, participants pick outcomes (e.g., heads or tails), while participants in our studies pick processes (e.g., what will determine a coin flip bet). Thus, the results in this paper suggest that having people choose among prediction process instead of outcomes could reduce gamblers fallacy.

Relatedly, this work suggests that the prediction process may moderate the strength of belief in the Law of Small Numbers, which is the belief that a sample randomly drawn from a population will represent that population in all essential characteristics (e.g., the distribution of

outcomes; see Tversky and Kahneman, 1971). Specifically, Study 1b suggests that people may believe that drawing a sample in a way that most closely matches the process that generates event outcomes is most likely to produce a sample that represents the population.

Limitations & Future Directions

There are several promising directions for future work on replicativeness. For example, future research could explore how much heterogeneity there is in the preference for replicative prediction algorithms, and what individual differences predict this preference. We proposed that one of the reasons why people like replicative algorithms is that they believe replicative algorithms will perform well. In contrast to this intuition, people who readily engage in statistical thinking might easily pick up on the fact that a replicative algorithm may perform poorly because it is based on a small sample, or doesn't always pick the optimal answer. Therefore, future studies could test whether people's statistical training, propensity for cognitive reflection (Frederick 2005), and other individual differences predict people's preference for replicative prediction algorithms.

Future studies could also further examine how people form their perceptions of replicativeness. We have proposed that these judgments may work similarly to judgments of representativeness in that the reflection of "essential properties" may be important (Tversky and Kahneman 1982, p.85), but more work is needed to understand exactly how people make these judgments. Relatedly, future work could investigate whether there are dimensions of prediction algorithms that are more influential than others when people are forming judgments of replicativeness. Finally, future work should investigate whether people's preference for replicative algorithms is similar to what we have found in other domains that we did not test.

Conclusion

A famous proverb states that “experience is the best teacher.” Although experiences may teach us valuable lessons about making accurate judgments and predictions, relying on experiences alone is a flawed philosophy in random domains because outcomes can differ substantially in similar circumstances. In spite of this, people’s preferences for prediction by replication suggest they often prefer to rely directly on experience to make predictions, assuming that the same process will lead to the same outcome. Our research raises awareness of this tendency both to build understanding of how people select prediction algorithms, and provide a tool that practitioners can use to promote more productive means of prediction when experience alone fails.

Appendix 1: Tables

Table 1.1: Stimuli from Study 2b, Chapter 1

Focal Items	**Comparison Items**		
	Close	**Intermediate**	**Far**
Chocolate	Pizza	Liquor	Toys
Juice	Milk	Pet Food	Jeans
Ski trips	Weekend Getaway	Holiday purchases	Medicine
Jeans	Shoes	Skates	Meats
Microwave	Pots and pans	Laptop	Weekend Getaway

Notes. –The norming ratings for each item is in Web Appendix C.

Table 1.2: Stimuli from Study 3a and 3b, Chapter 1

Focal	***Comparison***		
	Close	**Mid**	**Far**
gourmet chocolate truffles ***meats***	wine	gas	toys
sunglasses ***jeans***	backpack	lamp	restaurant gift cards
potted plants ***lamp***	pots and pans	books	sunglasses
diapers ***toilet paper***	detergents	stationary	pizza
restaurant gift cards ***movie tickets***	streaming services	gourmet chocolate truffles	pet food

Notes. – Red indicates that the focal item is typical of its category.

Table 1.3: Product Sets from Study 4, Chapter 1

Comparison	***Focal***	
	Close	**Far**
Detergents	Toilet Tissue (N = 2,082,158)	Coffee (N = 2,173,191)
Pizza	Coffee (N = 1,547,102)	Toilet Tissue (N = 1,152,999)
Detergents	Disposable Diapers (N = 165,649)	Pet Food (N = 44,980)
Pizza	Pet Food (N = 28,917)	Disposable Diapers (N =136,821)

Notes. – The number below each focal item indicate the number of trips that were used in the analysis.

Table 2.1: Recommendation Description from Study 2c, Chapter 2

Description	Wording
Baseline	"We collect data from our million + clients and use exclusive technology to narrow down clothing options that you will like."
Baseline+	"We identify your dimensions and stylistic preferences, and we apply those features with exclusive technology and data from our million + clients to narrow down clothing options that you will like."
Similar Features	"We identify your dimensions and stylistic preferences, and we use data from our million+ clients to find those who like the similar features and recommend items that they like."
Feature Combined	"We use exclusive technology to identify your dimensions and stylistic preferences and we use data from our million + clients to identify how these features can be best combined to generate options that you will like."
Specific Taste	"We collect data from our million + clients and use exclusive technology to narrow down clothing options. Specifically, we find the clothes that you will like given your dimensions, your stylistic preferences, and our knowledge about how to combine these features"
CF	"We collect data from our million + clients and use exclusive technology to narrow down clothing options. Specifically, we have built a system that picks clothes you will like based on how others with your dimensions and stylistic preferences like them"
CF + System	"We collect data from our million + clients and use exclusive technology to narrow down clothing options. Specifically, we have built a system that predicts what clothes people will like given their dimensions and stylistic preferences and we apply it to your specific dimensions and stylistic preferences."
Input	"We collect data from our million + clients and use exclusive technology to narrow down clothing options. Specifically, we input your dimensions and stylistic preferences into a system that predicts what clothes people will like given their dimensions and stylistic preferences."

Table 2.2: Dimension Wording from Study 2c, Chapter 2

Dimension	Wordings
Replicative 1:	This company's methodology shares similarities with how I choose clothes for myself. (1 – strongly disagree; 5 – strongly agree)
Replicative 2:	The way I pick out my clothes is reflected well in this company's methodology. (1 – strongly disagree; 5 – strongly agree)
Unique:	How unique is the method that the company is using? (1 – Not at all; 5 – Extremely)
Expensive:	How expensive do you think the service is? (1 – Not at all; 5 – Extremely)
Understanding:	How well do you understand how the method picks out clothes for you? (1 – Very Little; 5 – Completely)

Table 2.3: Conditions from Study 3, Chapter 2

	Accuracy	
Replicativeness	Target more accurate	Same performance
Target more replicative	Linear bonus. Alternative algorithm predicts 6.	Bonus only for perfect predictions. Alternative algorithm predicts 6.
Target less replicative	Linear bonus. Alternative algorithm rolls a uniform die (1, 2, 3, 4, 5, 6, 7)	Bonus only for perfect predictions. Alternative algorithm rolls a uniform die (1, 2, 3, 4, 5, 6, 7)

Table 2.4: Regression from Study 3, Chapter 2

Predictors	Coefficient (SE)
Choice Number	-.009*
	(.005)
More Accurate	.383***
	(.142)
More Replicative	.381***
	(.139)
Choice Number*More Accurate	.016***
	(0.006)
Choice Number*More Replicative	.002
	(0.006)
More Accurate*More Replicative	-0.004
	(.199)
Intercept	0.012
	(0.098)
df	795

Note: *p<0.1; **p<0.05; ***p<0.01

Appendix 2: Figures

Figure 1.1: Possible Representation of Expenditures, Chapter 1

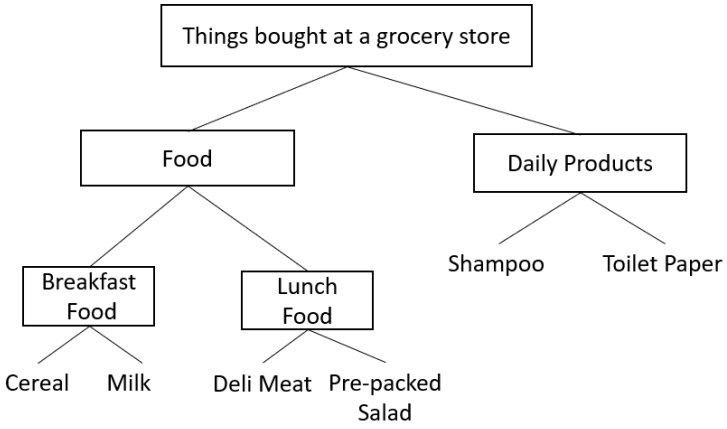
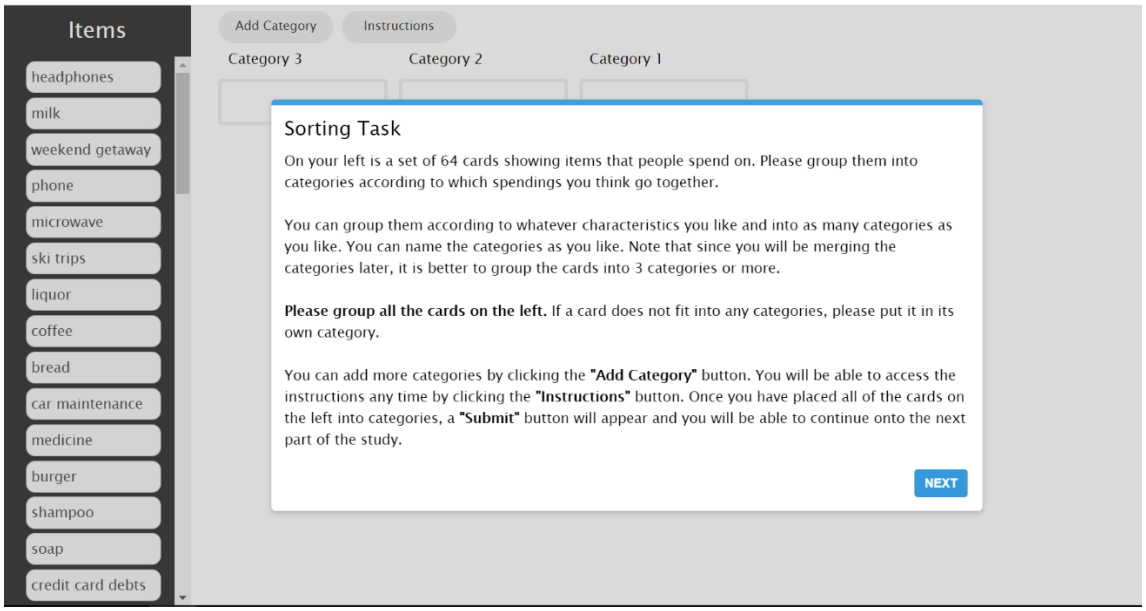
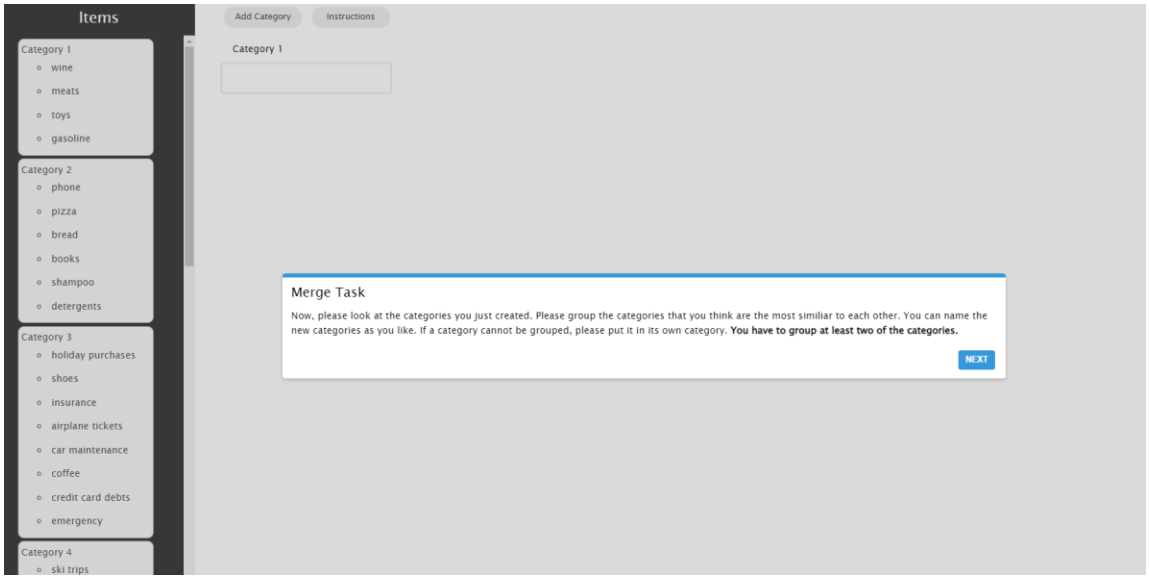
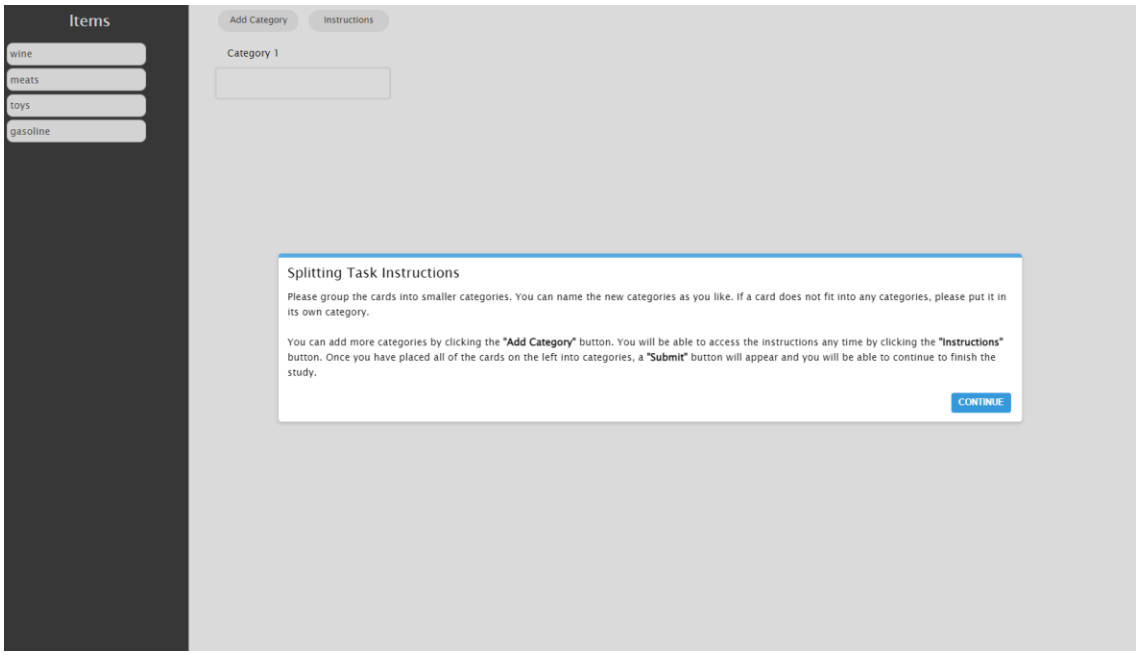


Figure 1.2: Successive Pile Sort Interface, Chapter 1

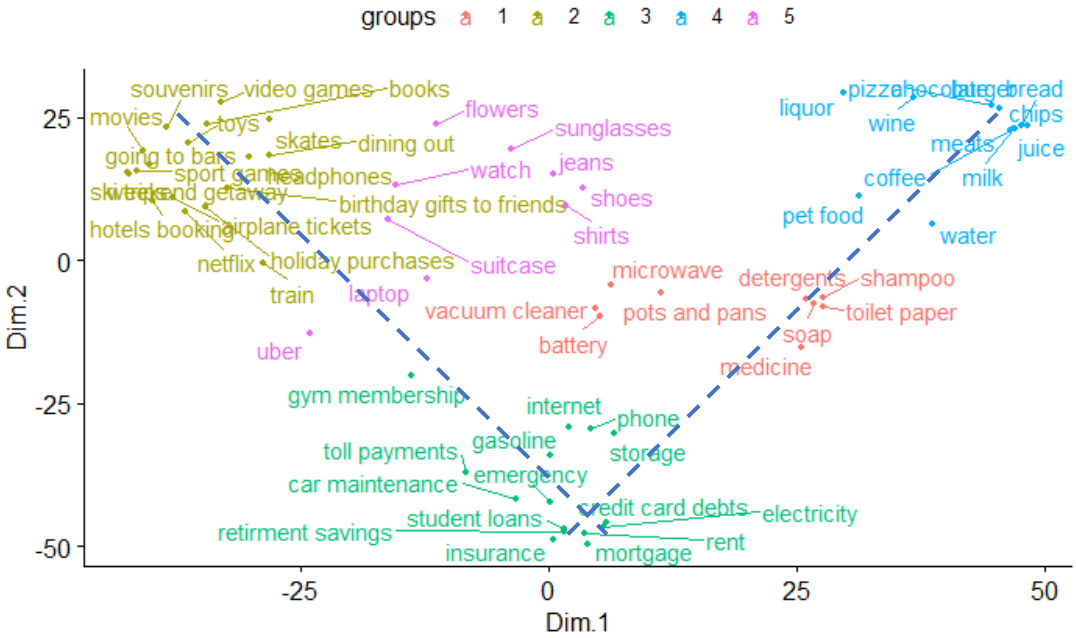






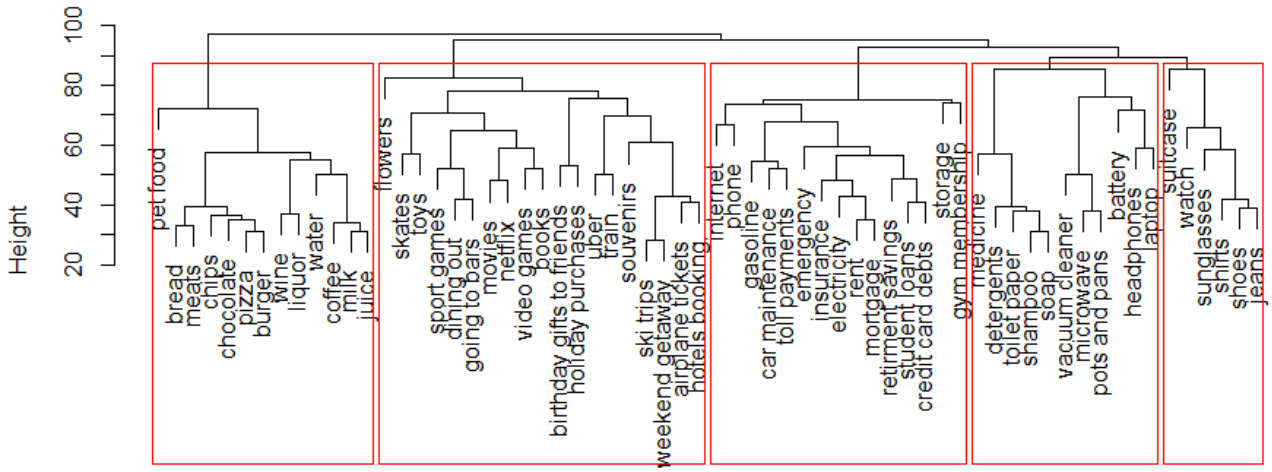
Note. – Interface for the successive pile-sort task. Top: Initial sort interface and instructions. Middle: Merging interface and instructions. Bottom: Splitting interface and instructions.

Figure 1.3: Multidimensional Scaling Reduction with Clustered Groups from Study 1, Chapter 1



Note. – Multidimensional Scaling reduces the space of products to two dimensions, but the two dimensions need not be the X and Y axis. Any orthogonal directions can be interpreted as the dimensions (Cox and Cox 2008). Dashed lines represent one interpretation. The line sloping down may represent how hedonic a product is (Hirschman and Holcrook 1982) while the line sloping up may represent the amount of spending or the frequency of purchasing.

Figure 1.4: Dendrogram of the Products on Aggregate Level from Study 1, Chapter 1



Note. – The red brackets represent the aggregate level categorizations. This grouping has a slight discrepancy with the MDS clustering because MDS maps the expenditures on two dimensions while the dendrogram only uses the aggregate distance matrix.

Figure 1.5: Spending Adjustment on Comparison Items from Study 2a, Chapter 1

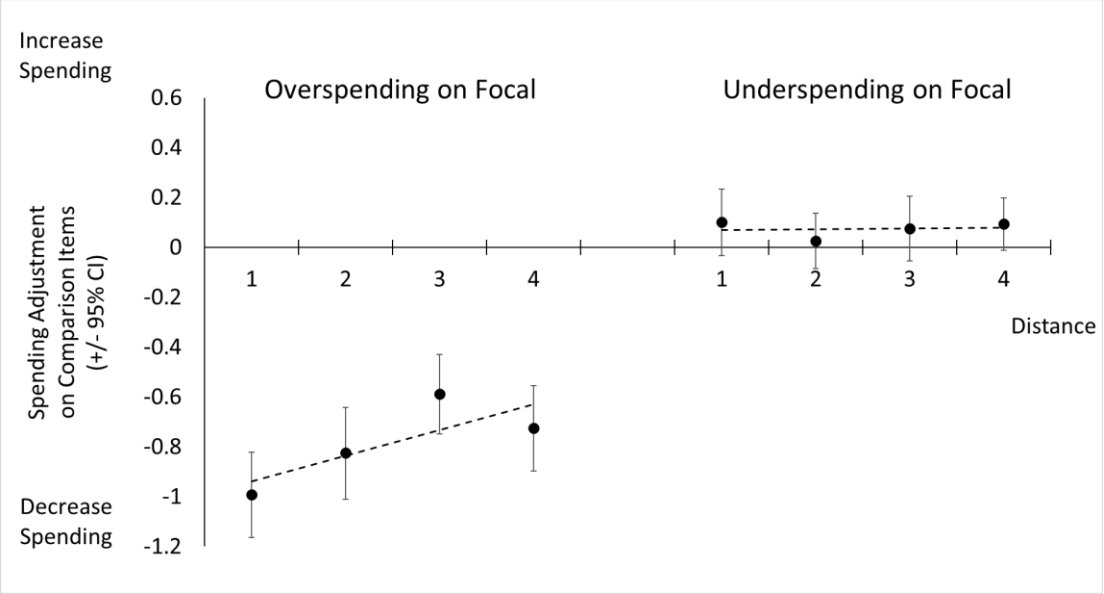


Figure 1.6: Spending Adjustment on Comparison Items from Study 2b, Chapter 1

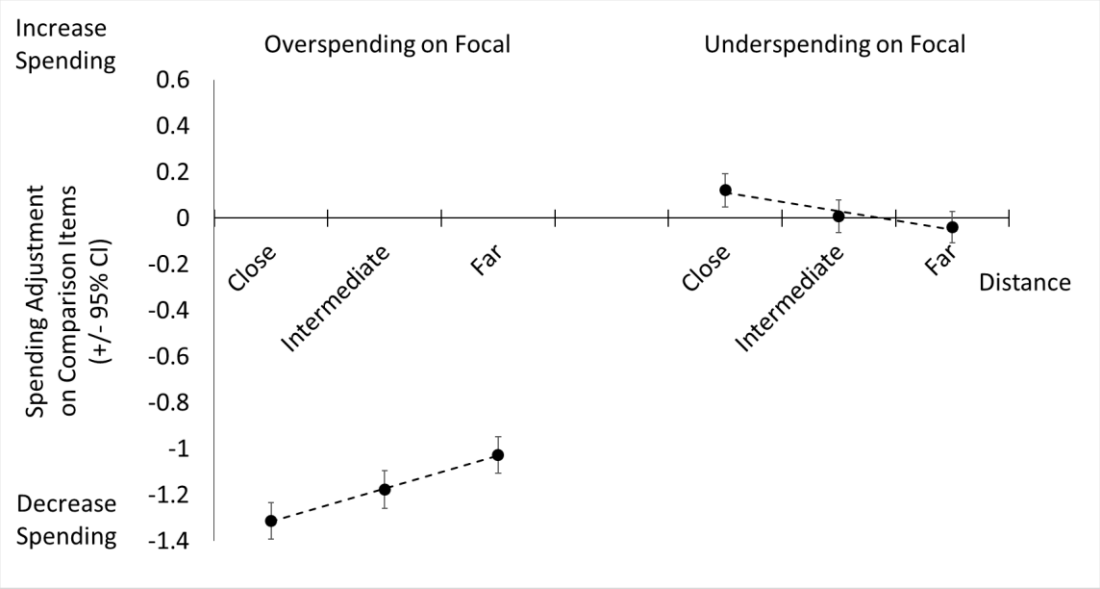


Figure 1.7: Raw Ratings and Estimated Marginal Means of Study 2c, Chapter 1

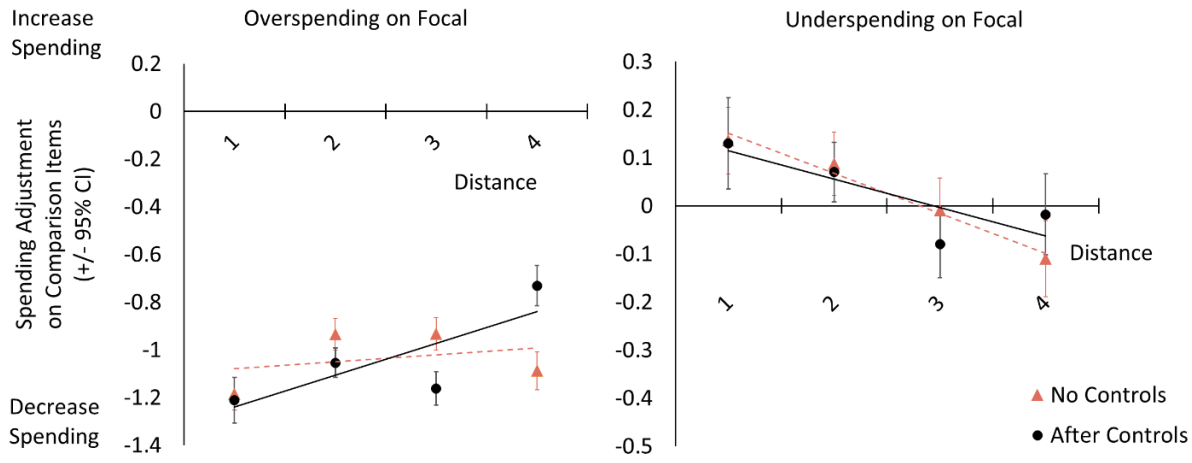
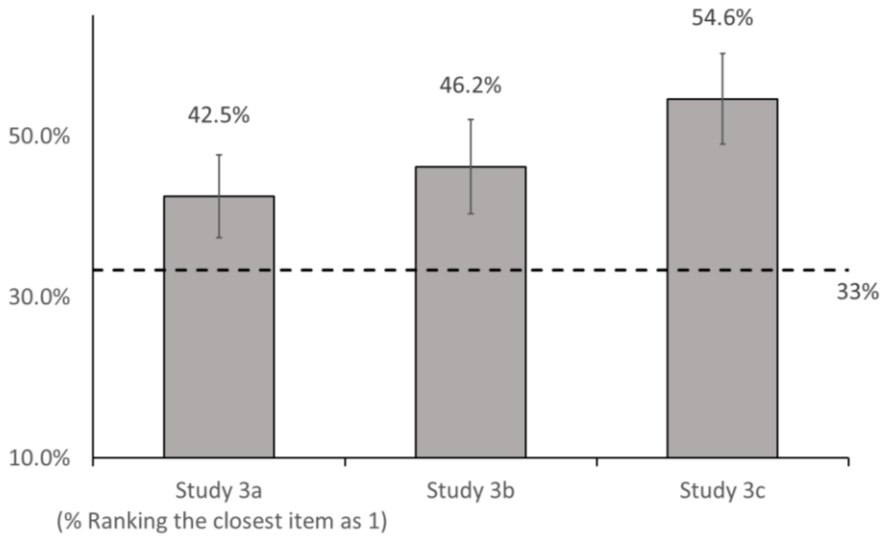


Figure 1.8: Proportions Choosing the Closest Comparison Product from Study 3, Chapter 1



Note. - (Left) Proportion of closest comparison products ranked 1st to apply promotions to in study 3a. (Middle and Right) Proportions choosing the closest comparison product to apply promotions to in study 3b and 3c. Horizontal line indicates the expected proportion (33%) if participants were choosing at random.

Figure 1.9: Regression Coefficients by Year on Close Focal vs Far Focal from Study 4, Chapter 1

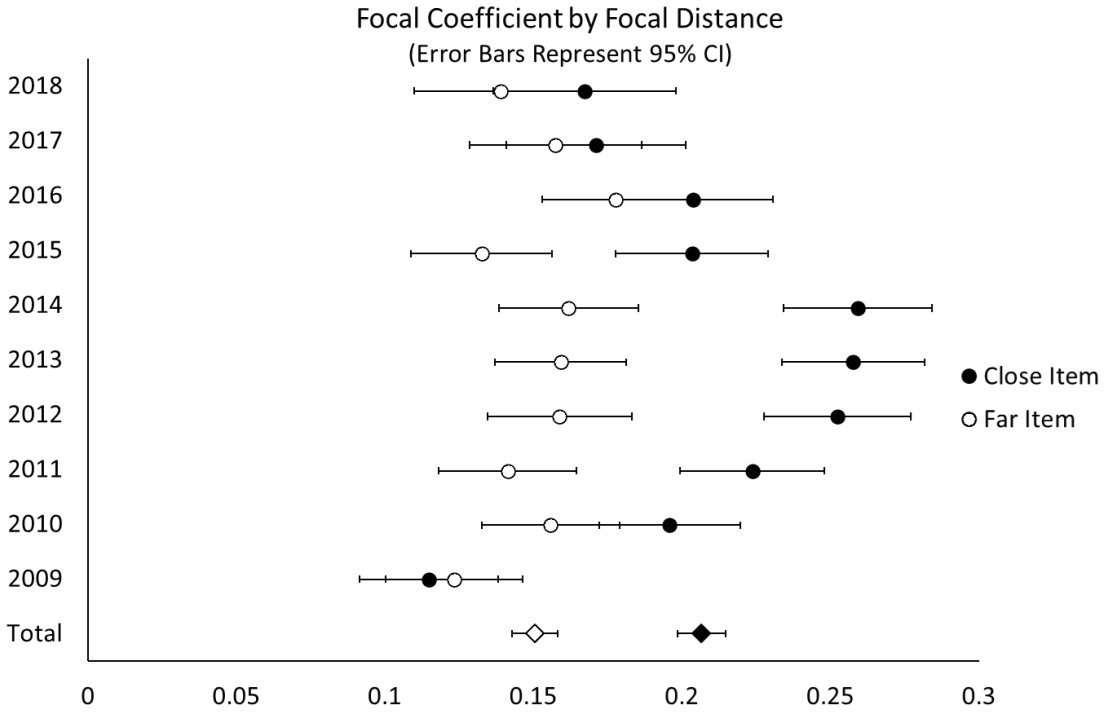
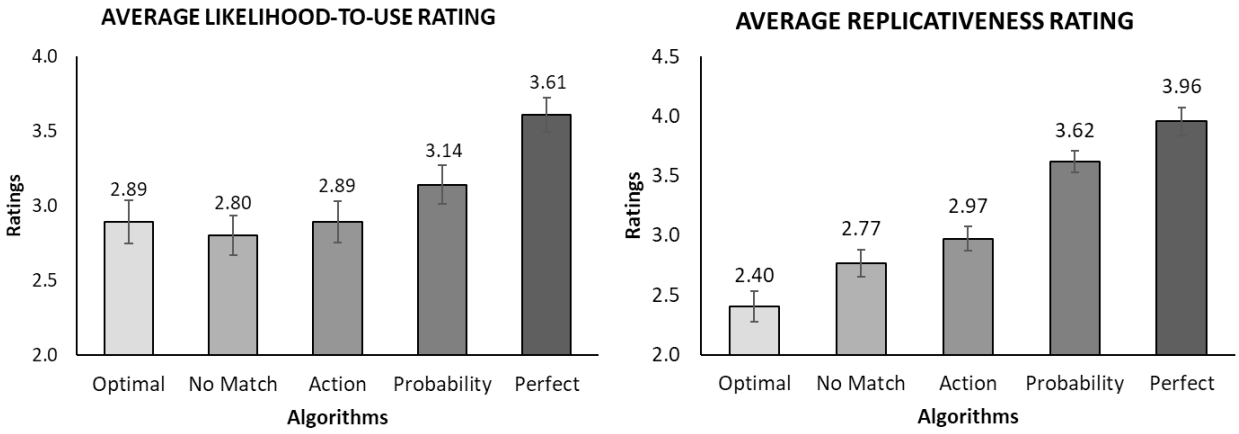
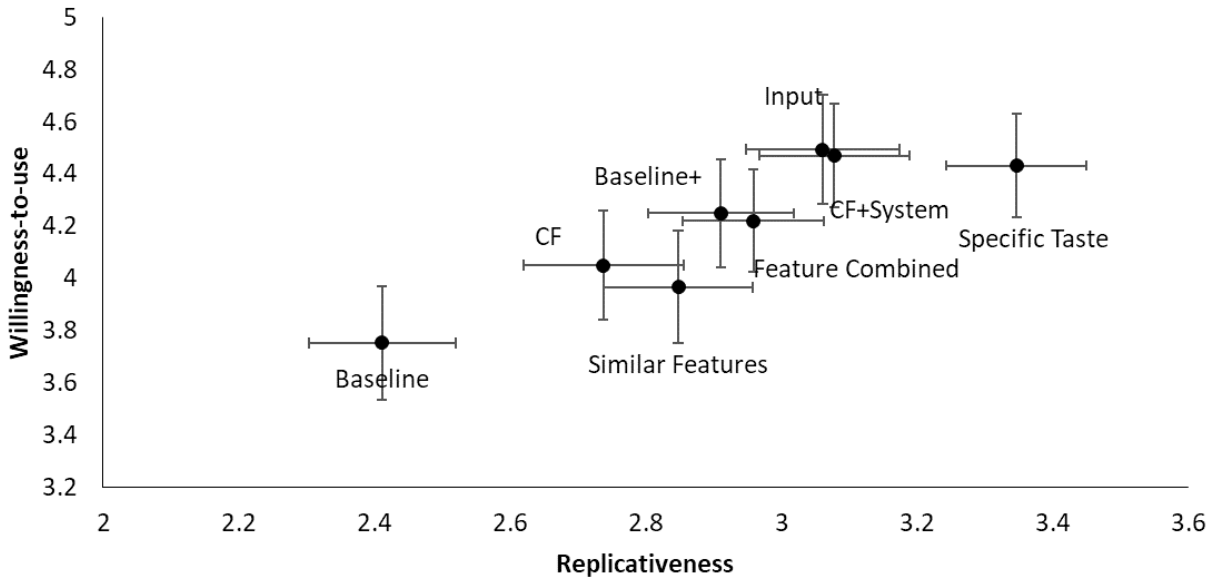


Figure 2.1: Likelihood-to-Use and Replicativeness Rating from Study 1b, Chapter 2



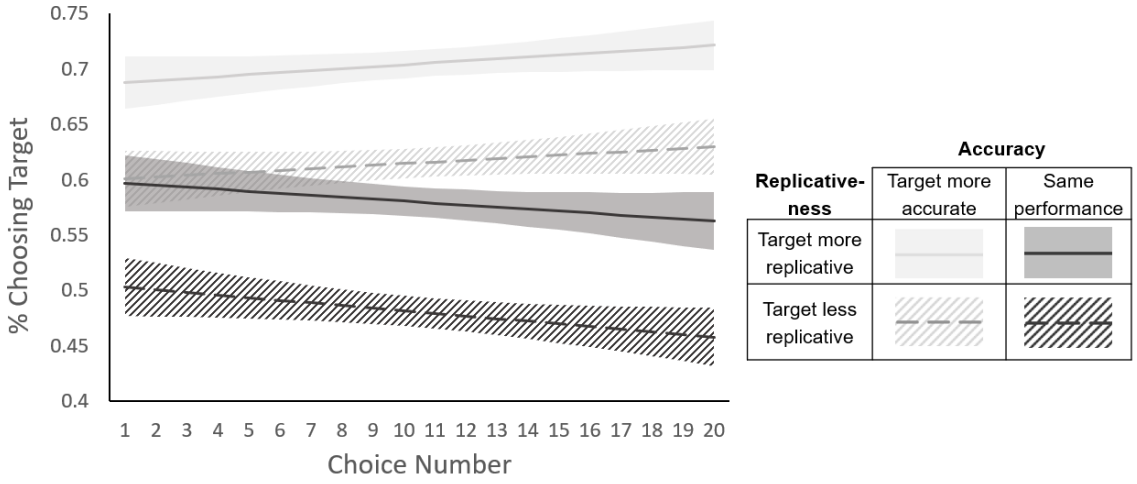
Note. - Average likelihood of using each of the algorithm in Study 1b (left) and average replicativeness rating (right). Error bars represent 95% confidence intervals.

Figure 2.2: Replicativeness & Willingness-to-Use Ratings from Study 2c, Chapter 2



Note. - Correlation between replicativeness rating (x-axis) and willingness-to-use rating (y-axis) in Study 2c. Vertical error bars represent 95% confidence interval for willingness to use, horizontal error bars represent 95% confidence interval for replicativeness.

Figure 2.3: Percentage Choosing the Target Algorithm from Study 3, Chapter 2



Note. - Model estimated percentage of participants choosing the target algorithm in each condition of Study 3. Ribbons indicate upper and lower bounds of the 95% interval.

References

Aaker, David A. and Kevin L. Keller (1990), "Consumer Evaluations of Brand Extensions," *Journal of Marketing*, 54 (January), 27–41.

Alba, Joseph W., and J. Wesley Hutchinson (1987), "Dimensions of Consumer Expertise," *Journal of Consumer Research*, 13(4), 411-54.

Antonides, Gerrit, I. Manon De Groot, and W. Fred Van Raaij (2011), "Mental Budgeting and the Management of Household Finance," *Journal of Economic Psychology* 32(4), 546-55.

Arkes, Hal R., Cynthia A. Joyner, Mark V. Pezzo, Jane Gradwohl Nash, Karen Siegel-Jacobs, and Eric Stone (1994), "The Psychology of Windfall Gains," *Organization Behavior and Human Decision Processes*, 59 (3), 331 – 47.

Atran, Scott, Douglas Medin, Norbert Ross, Elizabeth Lynch, John Coley, Edilberto Ucan Ek, and Valentina Vapnarsky (1999), "Folkecology and Commons Management in the Maya Lowlands," *Proceedings of the National Academy of Sciences*, 96(13), 7598-603.

Ayton, Peter, and Ilan Fischer (2004), "The Hot Hand Fallacy and the Gambler's Fallacy: Two Faces of Subjective Randomness?" *Memory & Cognition*, 32(8), 1369-378.

Balzer, William K., Michael E. Doherty, and Raymond O'Connor, Jr. (1989), "Effects of Cognitive Feedback on Performance," *Psychological Bulletin*, 106(3), 410-33.

Bandura, Albert, Dorothea Ross, and Sheila A. Ross (1961), "Transmission of Aggression Through Imitation of Aggressive Models," *The Journal of Abnormal and Social Psychology*, 63(November), 575-82.

Barsalou, Lawrence W. (1983), "Ad Hoc Categories," *Memory and Cognition*, 11 (3), 211–27.

———. (1985), “Ideals, Central Tendency, and Frequency of Instantiation as Determinants of Graded Structure in Categories,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 11 (4), 629–54.

Bartels, Daniel M., and Eric J. Johnson (2015), "Connecting Cognition and Consumer Choice," *Cognition*, 135, 47-51.

Bartels, Daniel M., and Eric J. Johnson (2015), “Connecting Cognition and Consumer Choice,” *Cognition,* 135, 47-51.

Berlin, Brent (1992), *Ethnobiological Classification: Principles of Categorization of Plants and Animals in Traditional Societies*. Princeton University Press.

Boster, James (1994), “The Successive Pile Sort,” *Cultural Anthropology Methods*, 6(2), 11-12.

Bottomley, Paul A., and Stephen JS Holden (2001), “Do We Really Know How Consumers Evaluate Brand Extensions? Empirical Generalizations Based on Secondary Analysis of Eight Studies,” *Journal of Marketing Research*, 38(4), 494-500.

Burger, Jerry M., Nicole Messian, Shebani Patel, Alicia Del Prado, and Carmen Anderson (2004), “What a Coincidence! The Effects of Incidental Similarity on Compliance,” *Personality and Social Psychology Bulletin,* 30(January), 35-43.

Castelo, Noah, Maarten W. Bos, and Donald Lehmann (2019), “Let the Machine Decide: When Consumers Trust or Distrust Algorithms.” *NIM Marketing Intelligence Review,* 11(2), 24-29.

Cheema, Amar and Dilip Soman (2006), “Malleable Mental Accounting: The Effect of Flexibility on the Justification of Attractive Spending and Consumption Decisions,” *Journal of Consumer Psychology,* 16(1), 33-44.

Chintagunta, Pradeep K., and Sudeep Haldar (1998), “Investigating Purchase Timing Behavior in Two Related Product Categories,” *Journal of Marketing Research*, 35(1), 43-53.

Clancy, Kevin, John Bartolomeo, David Richardson, and Charles Wellford (1981), “Sentence Decisionmaking: The Logic of Sentence Decisions and the Extent and Sources of Sentence Disparity,” *The Journal of Criminal Law and Criminology*, 72(2): 524-54.

Cohen, Joel B., and Kunal Basu (1987), “Alternative models of categorization: Toward a contingent processing framework,” *Journal of Consumer Research*, 13(4), 455-72.

Collins, Allan M., and Elizabeth F. Loftus (1975), “A Spreading-Activation Theory of Semantic Processing,” *Psychological Review*, 82(6), 407-28.

Collins, Allan M., and M. Ross Quillian (1969), “Retrieval Time from Semantic Memory.” *Journal of Verbal Learning and Verbal Behavior*, 8, 240-47.

Colson, Eric, Brian Coffey, Tarek Rached, and Liz Cruz (2021), “Stitch Fix Algorithms Tour” (accessed December 21, 2021) https://algorithms-tour.stitchfix.com/.

Cox, Michael AA, and Trevor F. Cox (2008), “Multidimensional Scaling.” In *Handbook of Data Visualization*, Springer, Berlin, Heidelberg, 315-47.

Dawes, Robyn M., David Faust, and Paul E. Meehl (1989), "Clinical Versus Actuarial Judgment." *Science*, 243(4899), 1668-1674.

Deaton, Angus, and John Muellbauer (1980), *Economics and Consumer Behavior*, Cambridge University Press.

Dietvorst, Berkeley J., and Daniel M. Bartels (2021), “Consumers Object to Algorithms Making Morally Relevant Tradeoffs Because of Algorithms’ Consequentialist Decision Strategies,” *Journal of Consumer Psychology*, DOI: 10.1002/jcpy.1266.

Dietvorst, Berkeley J., and Soaham Bharti (2020), "People Reject Algorithms in Uncertain Decision Domains Because They Have Diminishing Sensitivity to Forecasting Error," *Psychological Science*, 31(10), 1302-314.

Dietvorst, Berkeley J., Joseph P. Simmons, and Cade Massey (2015), "Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err," *Journal of Experimental Psychology: General*, *144*(February), 114-26.

Einhorn, Hillel J (1986), "Accepting Error to Make Less Error," *Journal of Personality Assessment*, 50(3), 387-95.

Estes, William K., and J. H. Straughan (1954), "Analysis of A Verbal Conditioning Situation in Terms of Statistical Learning Theory," *Journal of Experimental Psychology*, *47*(April), 225-34.

Felcher, E. Marla, Prashant Malaviya, and Ann L. McGill (2001), "The Role of Taxonomic and Goal‐Derived Product Categorization in, Within, and Across Category Judgments," *Psychology & Marketing*, 18(8), 865-87.

Fildes, Robert, and Paul Goodwin (2007), "Against Your Better Judgment? How Organizations Can Improve Their Use of Management Judgment in Forecasting," *Interfaces*, 37(6), 570-76.

Fox, Craig R., and Gülden Ülkümen (2011), "Distinguishing Two Dimensions of Uncertainty," *Perspectives on Thinking, Judging, and Decision Making*, 21-35.

Frederick, Shane (2005), "Cognitive Reflection and Decision Making," *Journal of Economic Perspectives*, 19(4), 25-42.

Gallagher, James (2017), "Artificial Intelligence as Good as Cancer Doctors," *BBC News*, January 26, https://www.bbc.com/news/health-38717928.

Hastie, Reid (2015), “Causal Thinking in Judgments,” *The Wiley Blackwell Handbook of Judgment and Decision Making*, 2, 590-628.

Hastings, Justine S. and Jesse M. Shapiro (2013), “Fungibility and Consumer Choice: Evidence from Commodity Price Shocks”, *The Quarterly Journal of Economics*, 128(4), 1449-498.

Hastings, Justine, and Jesse M. Shapiro (2018), “How Are SNAP Benefits Spent? Evidence from a Retail Panel.” *American Economic Review*, 108(12), 3493-540.

Hauser, John R., and Frank S. Koppelman (1979), “Alternative Perceptual Mapping Techniques: Relative Accuracy and Usefulness,” *Journal of Marketing Research*, 16(4), 495-506.

Heath, Chip and Jack B. Soll (1996), “Mental Budgeting and Consumer Decisions”, *Journal of Consumer Research*, 23(1), 40-52.

Henderson, Pamela W., and Robert A. Peterson (1992), “Mental Accounting and Categorization,” *Organizational Behavior and Human Decision Processes*, 51 (2), 92–117.

Hirschman, Elizabeth C., and Morris B. Holbrook (1982), “Hedonic Consumption: Emerging Concepts, Methods and Propositions,” *Journal of Marketing*, 46(3), 92-101.

Hoehl, Stefanie, Stefanie Keupp, Hanna Schleihauf, Nicola McGuigan, David Buttelmann, and Andrew Whiten (2019). “‘Over-imitation’: A Review and Appraisal of a Decade of Research,” *Developmental Review*, 51, 90-108.

Horner, Victoria, and Andrew Whiten (2005), “Causal Knowledge and Imitation/Emulation Switching in Chimpanzees (Pan Troglodytes) and Children (Homo Sapiens),” *Animal Cognition*, 8(3), 164-81.

Huber, Joel, John W. Payne, and Christopher Puto (1982), "Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis," *Journal of Consumer Research*, 9(1), 90-8.

Jacko, Julie A., and Gavriel Salvendy (1996), “Hierarchical Menu Design: Breadth, Depth, and Task Complexity,” *Perceptual and Motor Skills,* 82(3), 1187-201.

Jiang, Lan, Joandrea Hoegg, Darren W. Dahl, and Amitava Chattopadhyay (2010), “The Persuasive Role of Incidental Similarity on Attitudes and Purchase Intentions in A Sales Context,” *Journal of Consumer Research,* 36(5), 778-91.

Kahneman, Daniel and Amos Tversky (1984), “Choices, Values, and Frames,” *American Psychologist*, 39, 341 – 50.

Kahneman, Daniel, and Amos Tversky (1972), "Subjective Probability: A Judgment of Representativeness," *Cognitive Psychology*, 3(3), 430-54.

Kahneman, Daniel, and Shane Frederick (2002), “Representativeness Revisited: Attribute Substitution in Intuitive Judgment,” *Heuristics and Biases: The Psychology of Intuitive Judgment,* 49-81.

Kahneman, Daniel, Olivier Sibony, and Cass R. Sunstein (2021), *Noise: a Flaw in Human Judgment*. Little, Brown.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan (2018), "Human Decisions and Machine Predictions," *The Quarterly Journal of Economics*, 133(1), 237-293.

Krishnamurthy, Parthasarathy, and Sonja Prokopec (2010), “Resisting That Triple-Chocolate Cake: Mental Budgets and Self-Control,” *Journal of Consumer Research*, 37(1), 68-79.

Kruskal, Joseph B. (1978), *Multidimensional Scaling*. No. 11. Sage.

Lagnado, David A., and Steven A. Sloman (2006), “Time as a Guide to Cause,” *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 32(3), 451-60.

Leachman, Sancy A., and Glenn Merlino (2017), "The Final Frontier in Cancer Diagnosis," *Nature*, 542(7639), 36-8.

Levav, Jonathan and A. Peter McGraw (2009), "Emotional Accounting: How Feelings About Money Influence Consumer Choice," *Journal of Marketing Research*, 46, 66–80.

Linville, Patricia W., and Gregory W. Fischer (1991), "Preferences for Separating or Combining Events," *Journal of Personality and Social Psychology*, 60(1), 5-23.

Logg, Jennifer M., Julia A. Minson, and Don A. Moore (2019), "Algorithm Appreciation: People Prefer Algorithmic to Human Judgment," *Organizational Behavior and Human Decision Processes*, 151, 90-103.

Lohr, Steve (2016), "IBM Is Counting on Its Bet on Watson, and Paying Big Money for It," *New York Times*, 17(Oct), https://www.nytimes.com/2016/10/17/technology/ibm-is-counting-on-its-bet-on-watson-and-paying-big-money-for-it.html.

Longoni, Chiara, Andrea Bonezzi, and Carey K. Morewedge (2019), "Resistance to Medical Artificial Intelligence," *Journal of Consumer Research*, 46(4), 629-650.

Lopez, Alejandro, Scott Atran, John D. Coley, Douglas L. Medin, and Edward E. Smith (1997), "The Tree of Life: Universal and Cultural Features of Folkbiological Taxonomies and Inductions," *Cognitive Psychology*, 32, 251-95.

Lynch Jr, John G., Richard G. Netemeyer, Stephen A. Spiller, and Alessandra Zammit (2010), "A Generalizable Scale of Propensity to Plan: The Long and the Short of Planning for Time and for Money," *Journal of Consumer Research*, 37(1), 108-28.

Lynch, Peter (2008), "The Origins of Computer Weather Prediction and Climate Modeling," *Journal of Computational Physics*, 227(7), 3431-444.

Manchanda, Puneet, Asim Ansari, and Sunil Gupta (1999), "The "Shopping Basket": a Model for Multicategory Purchase Incidence Decisions," *Marketing Science,* 18(2), 95-114.

Markman, Arthur B. (2002), "Knowledge Representation."

Markman, Arthur B., C. Miguel Brendl, and Kyungil Kim (2007), "Preference and the Specificity of Goals," Emotion 7(3), 680-84.

McCracken, J., C. Osterhout, and James F. Voss. (1962), "Effects of Instructions in Probability Learning," *Journal of Experimental Psychology,* 64(3), 267.

McCullagh, Peter (1980), "Regression Models for Ordinal Data," *Journal of the Royal Statistical Society: Series B (Methodological),* 42(2), 109-127.

McGuire, William J. (1978), "An Information-Processing Model of Advertising Effectiveness," in *Behavioral and Management Sciences in Marketing,* ed. Harry L. Davis and Alvin J. Silk, New York: Wiley.

Medin, Douglas L., and Scott Atran (2004), "The Native Mind: Biological Categorization and Reasoning in Development and Across Cultures," *Psychological Review,* 111(4), 960-83.

Medin, Douglas L., Elizabeth B. Lynch, John D. Coley, and Scott Atran (1997), "Categorization and Reasoning Among Tree Experts: Do All Roads Lead to Rome?" *Cognitive Psychology,* 32, 49-96.

Meehl, Paul E (1954),"Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence," University of Minnesota Press. https://doi.org/10.1037/11281-000.

Miller, Dwight P (1981), "The Depth/Breadth Tradeoff in Hierarchical Computer Menus," *Proceedings of the Human Factors Society Annual Meeting,* 25(1), 296-300.

Moreau, C. Page, Arthur B. Markman, and Donald R. Lehmann (2001), ""What is It?" Categorization Flexibility and Consumers' Responses to Really New Products," *Journal of Consumer Research*, 27(4), 489-98.

Moreau, C. Page, Donald R. Lehmann, and Arthur B. Markman (2001), "Entrenched Knowledge Structures and Consumer Response to New Products," *Journal of Marketing Research*, 38(1),14-29.

Morewedge, Carey K., Leif Holtzman, and Nicholas Epley (2007), "Unfixed Resources: Perceived Costs, Consumption, and the Accessible Account Effect," *Journal of Consumer Research,* 34(4), 459-67.

Murphy, Gregory L. (2002), *The Big Book of Concepts*, Cambridge, MA: MIT Press.

Nedungadi, Prakash (1990), "Recall and Consumer Consideration Sets: Influencing Choice Without Altering Brand Evaluations," *Journal of Consumer Research,* 17(3), 263-76.

Nedungadi, Prakash (1990), "Recall and Consumer Consideration Sets: Influencing Choice Without Altering Brand Evaluations," *Journal of Consumer Research*, 17(3), 263-76.

Osherson, Daniel N., Edward E. Smith, Ormond Wilkie, Alejandro Lopez, and Eldar Shafir (1990), "Category-Based Induction," *Psychological Review*, 97(2), 185-200.

Pandora Media (2021), "Music Genome Project" (accessed December 21, 2021), https://www.pandora.com/about/mgp

Parker, Jeffrey R., Donald R. Lehmann, Kevin Lane Keller, and Martin G. Schleicher (2018), "Building a Multi-Category Brand: When Should Distant Brand Extensions Be Introduced?" *Journal of the Academy of Marketing Science*, 46(2), 300-16.

Payne, John W., John William Payne, James R. Bettman, and Eric J. Johnson (1993), T*he Adaptive Decision Maker*, Cambridge University Press.

Petty, Richard E., John T. Cacioppo, Richard E. Petty, and John T. Cacioppo (1986), *The Elaboration Likelihood Model of Persuasion*. Springer New York.

Randall, Robert A (1976), "How Tall is a Taxonomic Tree? Some Evidence for Dwarfism," *American Ethnologist*, 543-53.

Ratneshwar, S. and Allan D. Shocker (1991), "Substitution in Use and the Role of Usage Context in Product Category Structures," *Journal of Consumer Research*, 28 (August), 281–95.

Reinholtz, Nicholas, Daniel M. Bartels, and Jeffrey R. Parker (2015), "On the Mental Accounting of Restricted-Use Funds: How Gift Cards Change What People Purchase." *Journal of Consumer Research*, 42 (8), 596-614.

Rick, Scott I., Cynthia E. Cryder and George Loewenstein (2008), "Tightwads and Spendthrifts", *Journal of Consumer Research*, 34(6), 767-782.

Romney, A. Kimball, Susan C. Weller, and William H. Batchelder (1986), "Culture as Consensus: A Theory of Culture and Informant Accuracy," *American Anthropologist*, 88, 313-38.

Rosch, Eleanor, and Barbara Bloom Lloyd (1978), "Cognition and categorization," Lawrence Erlbaum Associates, Publishers.

Rosch, Eleanor, and Carolyn B. Mervis (1975), "Family Resemblances: Studies in the Internal Structure of Categories," *Cognitive Psychology*, 7(4), 573-605.

Rosch, Eleanor, Carolyn B. Mervis, Wayne D. Gray, David M. Johnson, and Penny Boyes-Braem (1976), "Basic Objects in Natural Categories," *Cognitive Psychology*, 8, 382–439.

Ross, B. H. (1997), "The use of categories affects classification," *Journal of Memory and Language*, 37, 240–267.

Schafer, J. Ben, Dan Frankowski, Jon Herlocker, and Shilad Sen (2007), “Collaborative Filtering Recommender Systems,” *The Adaptive Web,* 291-324.

Shanks, David R., Richard J. Tunney, and John D. McCarthy (2002), “A Re-Examination of Probability Matching and Rational Choice,” *Journal of Behavioral Decision Making*, 15(3), 233-50.

Shepard, Roger N. (1962), “The Analysis of Proximities: Multidimensional Scaling with An Unknown Distance Function. I,” *Psychometrika*, 27(2), 125-40.

Silver, Nate, Jay Boice, and Neil Payne (2021), “How our NBA Predictions Work” (accessed December 21, 2021), https://fivethirtyeight.com/methodology/how-our-nba-predictions-work/.

Sloman, S. A. (1998), “Categorical Inference is Not a Tree: The Myth of Inheritance Hierarchies,” *Cognitive Psychology*, 35, 1–33.

Smith, Edward E. (1978), “Theories of Semantic Memory,” *Handbook of Learning and Cognitive Processes,* 6, 1-56.

Solomon, Karen O., Douglas L. Medin, and Elizabeth Lynch (1999), “Concepts Do More Than Categorize,” *Trends in Cognitive Sciences,* 3(3), 99-105.

Soman, Dilip (2004), “Framing, Loss Aversion, and Mental Accounting,” *Blackwell Handbook of Judgment and Decision Making, ed. Koehler Derek J. Harvey Nigel, Oxford, U : Blackwell,* 379 – 98

Soman, Dilip, and Amar Cheema (2011) “Earmarking and Partitioning: Increasing Saving by Low-Income Households,” *Journal of Marketing Research*, 48, S14-S22.

Stewart, Neil, Nick Chater, and Gordon DA Brown (2006), "Decision by Sampling," *Cognitive Psychology,* 53(1), 1-26.

Sussman, Abigail B., and Adam L. Alter (2012), “The Exception is the Rule: Underestimating and Overspending on Exceptional Expenses.” *Journal of Consumer Research*, 39(4), 800-14.

Thaler, Richard H. (1985), “Mental Accounting and Consumer Choice,” *Marketing Science*, 27 (1), 15–25.

———. (1999). Mental Accounting Matters. *Journal of Behavioral Decision Making*, 12(3), 183-206.

———. (2015), *Misbehaving: The Making of Behavioral Economics*, W. W. Norton & Company.

Thaler, Richard H., and Eric J. Johnson (1990), “Gambling with the House Money and Trying to Break Even: The Effects of Prior Outcomes on Risky Choice,” *Management Science*, 36(6), 643-60.

Thaler, Richard H., and Hersh M. Shefrin (1981), “An Economic Theory of Self-Control.” *Journal of Political Economy*, 89(2), 392-406.

Trueblood, Jennifer S., Scott D. Brown, and Andrew Heathcote (2014), "The Multiattribute Linear Ballistic Accumulator Model of Context Effects in Multialternative Choice," *Psychological Review*, 121(2), 179-205.

Tversky, Amos, and Daniel Kahneman (1971), “Belief in The Law of Small Numbers,” *Psychological Bulletin*, 76(2), 105-10.

Tversky, Amos, and Daniel Kahneman (1974), "Judgment Under Uncertainty: Heuristics and Biases," *Science*, 185(4157), 1124-1131.

Tversky, Amos, and Daniel Kahneman (1982), "Judgments of and by Representativeness." *Judgment under Uncertainty: Heuristics and Biases*, 84-98.

Vrieze, Scott I., and William M. Grove (2009), "Survey on the Use of Clinical and Mechanical Prediction Methods in Clinical Psychology," *Professional Psychology: Research and Practice*, 40(5), 525-31.

Weiner, Bernard (1985), "'Spontaneous' causal thinking." *Psychological bulletin*, 97(1), 74-84.

Weller, Susan C (2007), "Cultural Consensus Theory: Applications and Frequently Asked Questions," *Field Methods*, 19(4), 339-68.

West, Richard F., and Keith E. Stanovich (2003), "Is Probability Matching Smart? Associations Between Probabilistic Choices and Cognitive Ability," *Memory & Cognition*, 31, 243-51.

Whiten, Andrew, Gillian Allan, Siobahn Devlin, Natalie Kseib, Nicola Raw, and Nicola McGuigan (2016), "Social Learning in The Real-World: 'Over-Imitation' Occurs in Both Children and Adults Unaware of Participation in An Experiment and Independently of Social Interaction," *PLoS One*, 11(7).

Woodside, Arch G., and J. William Davenport Jr. (1974), "The Effect of Salesman Similarity and Expertise on Consumer Purchasing Behavior," *Journal of Marketing Research*, 11(2), 198-202.

Yeomans, Michael, Anuj Shah, Sendhil Mullainathan, and Jon Kleinberg (2019), "Making Sense of Recommendations," *Journal of Behavioral Decision Making*, 32(4), 403-414.

Zhang, C. Yiwei, Abigail B. Sussman, Nathan Wang-Ly, and Jennifer Lyu (2020), "How Consumers Budget." Available at SSRN 3739543.