Journal of Experimental Psychology: General

ISSN: 0096-3445

2024, Vol. 153, No. 7, 1838–1860
https://doi.org/10.1037/xge0001607

Exploring Variability in Risk Taking With Large Language Models

Sudeep Bhatia
Department of Psychology, University of Pennsylvania

What are the sources of individual-level differences in risk taking, and how do they depend on the domain or situation in which the decision is being made? Psychologists currently answer such questions with psychometric methods, which analyze correlations across participant responses in survey data sets. In this article, we analyze the preferences that give rise to these correlations. Our approach uses (a) large language models (LLMs) to quantify everyday risky behaviors in terms of the attributes or reasons that may describe those behaviors, and (b) decision models to map these attributes and reasons onto participant responses. We show that LLM-based decision models can explain observed correlations between behaviors in terms of the reasons different behaviors elicit and explain observed correlations between individuals in terms of the weights different individuals place on reasons, thereby providing a decision theoretic foundation for psychometric findings. Since LLMs can generate quantitative representations for nearly any naturalistic decision, they can be used to make accurate out-of-sample predictions for hundreds of everyday behaviors, predict the reasons why people may or may not want to engage in these behaviors, and interpret these reasons in terms of core psychological constructs. Our approach has important theoretical and practical implications for the study of heterogeneity in everyday behavior.

Public Significance Statement
This article illustrates the value of large language models for understanding how people make everyday risky decisions, offering a novel approach to analyzing why individuals vary in their behaviors. By mapping out the reasons behind behaviors, it also sheds light on the psychological factors influencing decision making. This facilitates practical applications in predicting, interpreting, and improving real-world behaviors.

Keywords: decision making, risk, individual differences, computational models, artificial intelligence

The decisions that people make in everyday life determine their well-being, as well as the structure of our societies, governments, and economies. Understandably, considerable research in psychology has focused on understanding why people make the decisions they do, why different people may make different decisions, and why different domains and situations may elicit different decisions from the same individual (for reviews, see Edwards, 1961; Slovic et al., 1977; Weber & Johnson, 2009). Traditionally, much of this work falls within the psychometric tradition. Typically, researchers ask individuals to rate their tendencies to think or act in a certain way, and based on the correlational structure of the observed data, group individuals and items into coherent clusters, and study how these clusters differ based on situational, demographic, cultural, or personality variables.

The psychometric approach has made significant contributions to the study of decision making and has been widely adopted by researchers and practitioners working on risky behavior (e.g., Lauriola et al., 2007; Slovic, 1987; Weber et al., 2002), consumer behavior (e.g., Rick et al., 2008; Tian et al., 2001), social behavior (e.g., Graham et al., 2009; Rushton et al., 1981), as well as associated decision styles (e.g., Cacioppo & Petty, 1982; Epstein et al., 1996; Schwartz et al., 2002) and decision outcomes (e.g., Bruine de Bruin et al., 2007; Peters et al., 2006). However, by itself, it is unable to provide a complete account of the sources of variance in behavior. Psychometrics models the variance revealed by the data, and does not directly try to explain this variance in terms of cognitive and decision processes. Additionally, the psychometric approach relies exclusively on the statistical relationships within the collected data. Without knowing the reasons and motivations behind people's responses, it cannot anticipate how they will behave in novel circumstances or respond to new survey items for which ratings have not yet been collected.

A different line of research assumes that engaging in a particular behavior or making a particular decision involves different attributes or reasons, and that decision makers place weights on these attributes or reasons depending on their goals or desires. In this way, responses can be captured using a "decision model," which quantifies a behavior or decision in terms of its attribute values and quantifies the decision maker's preferences using attribute weights. Decision models have

This article was published Online First May 2, 2024.
Abigail Sussman served as action editor.
Sudeep Bhatia https://orcid.org/0000-0001-6068-684X
The material in this manuscript was previously presented at the 2023 Conference of the Society for Judgment and Decision Making. The funding for this study was received from the National Science Foundation (Grant SES-1847794). This is a sole-authored article and all work was done by Sudeep Bhatia. The design of the studies in this article was preregistered at https://osf.io/amves, and the code and data are available at https://osf.io/3y6ku.
Correspondence concerning this article should be addressed to Sudeep Bhatia, Department of Psychology, University of Pennsylvania, 425 South University Avenue, Philadelphia, PA 19104, United States. Email: bhatiasu@sas.upenn.edu

been able to explain and predict various phenomena observed in laboratory tasks in which the attribute structures of the decision alternatives are explicitly provided to the decision makers. These include risky (He, Analytis, & Bhatia, 2022; He, Zhao, & Bhatia, 2022; Kahneman & Tversky, 1979; Markowitz, 1959), intertemporal (He et al., 2023; Loewenstein & Prelec, 1992), social (Camerer, 2011; Fehr & Schmidt, 1999), managerial (Howard, 1988; Keeney & Raiffa, 1993), and consumer (Green & Srinivasan, 1990; J. Payne et al., 1991) choice. Decision models also relate choice behavior to core cognitive mechanisms like memory and attention, which determine how people sample and integrate attributes and reasons over the time course of the decision (Gigerenzer & Gaissmaier, 2011; Lee & Cummins, 2004; J. W. Payne et al., 1988; Roe et al., 2001; Weber et al., 2007; Zhao et al., 2022). For this reason, decision models provide a popular framework for describing choice in psychology, as well as closely related fields like behavioral economics, marketing, management, and cognitive neuroscience (see Bhatia et al., 2021; He, Analytis, & Bhatia, 2022; He, Zhao, & Bhatia, 2022 for recent reviews of existing models).

Importantly, decision models also provide a theoretically grounded explanation for item- and individual-level differences in behavior. According to the decision modeling approach, the similarity of responses across different items can be explained by the items' attribute structures: Items that have similar attributes will be judged similarly by the decision makers. Likewise, the similarity of responses across different individuals can be explained by individuals' preference functions: Individuals that assign similar weights to the attributes will make similar decisions.

Despite their appeal, decision models have limited applicability to real-world behaviors, where the attribute structures are often implicit and uncertain. In the laboratory, experimenters can manipulate one or two attributes and measure their effects on preferences, but in the real-world, common behaviors (such as investing in a speculative stock, mountain climbing, eating high cholesterol foods, shoplifting, and trusting a stranger with personal information) have many more complex reasons and attributes that influence them. Therefore, there is a gap between decision models of behavior and the actual behaviors that people exhibit in naturalistic contexts.

In this article we attempt to solve this problem using large language models (LLMs) (Brown et al., 2020; Devlin et al., 2018; Reimers & Gurevych, 2019; also see Mikolov et al., 2013). LLMs are artificial intelligence (AI) systems that can understand natural language after being trained on massive amounts of text data. They have shown remarkable capabilities in various natural language processing tasks, such as question answering, text summarization, content creation, and language translation, and are being widely used to model the mental representations that underlie human cognition and behavior (for reviews, see Bhatia & Aka, 2022; Bhatia et al., 2019). The reason for these successes is the ability of LLMs to quantify language by transforming words and sentences into numerical representations that capture their meanings. These numerical representations are used for prediction and language generation. Our goal is to use the representations obtained from LLMs to quantify the attributes that underlie everyday behaviors, and by doing so, apply the decision modeling approach to both predict how and understand why individuals vary in their behavior, and why different items elicit different behaviors from individuals.

We apply our approach to one of the most important topics in decision-making research: risk taking. Our empirical plan builds on a prominent existing scale for modeling domain and individual differences in everyday risky behavior, the Domain-Specific Risk Taking (DOSPERT) scale (Blais & Weber, 2006; Weber et al., 2002). DOSPERT measures people's tendencies to engage in several different types of common risky behaviors, which are presented to participants using short phrases and sentences. We present an expanded version of this scale to participants, and, with the use of LLM-derived attribute representations for the behaviors, attempt to both predict and understand people's responses, as well as the covariance structure in responses across items and individuals. We also use preference functions derived from people's choices to model the qualitative reasons that people use to explain their behavior. In this way, our approach attempts to combine decision modeling and psychometric research traditions, and by doing so, provide a comprehensive theoretical account of the sources of variance in everyday decision making.

Overview of Approach

Psychometrics

Psychometrics is central to the study of individual variability in several subdisciplines of psychology. Most relevant to this article is work that has applied psychometric techniques to study risky decision making. One of the most prominent instruments in this area, the DOSPERT, uses an inventory of 30 or 40 common behaviors sampled from five different domains: health, recreation, social behavior, ethical behavior, and financial decision making. It asks people to rate their propensity of committing each of the behaviors in the scale, and uses observed ratings to understand the differences between different individuals and groups, as well as between different items and domains (Blais & Weber, 2006; Weber et al., 2002).

Formally, the rating for behavior item i elicited from participant j can be written as r_{ij}. Subsequently, the list of all participant ratings for behavior i can be written as $\boldsymbol{r}_i^b = [r_{i1}, r_{i2}, \ldots r_{iM}]$ and the list of all behavior ratings by a participant j can be written as $\boldsymbol{r}_j^p = [r_{1j}, r_{2j}, \ldots r_{Nj}]$, where M and N are the number of participants and behaviors (items), respectively. With this notation, it is possible to quantify the similarity between two behaviors, i and i', in terms of their correlation across the individuals in a data set, $\mathrm{COR}(\boldsymbol{r}_i^b, \boldsymbol{r}_{i'}^b)$, and likewise quantify the similarity between two participants, j and j', in terms of their correlation across the behaviors in the data set, $\mathrm{COR}(\boldsymbol{r}_j^p, \boldsymbol{r}_{j'}^p)$. Psychometric methods often use these correlations to uncover latent factors or dimensions that describe the structure of variability in people's behaviors.

Researchers working with DOSPERT and related measures have found that, contrary to economic theories of decision making that posit a single general risk taking propensity, risk taking in everyday life varies significantly across domains (Dohmen et al., 2011; Frey et al., 2017; Highhouse et al., 2016; Weber et al., 2002). For example, the tendency to engage in a recreational risk (like mountain climbing) is strongly correlated with the tendency to engage in another recreational risk (like scuba diving), but not necessarily with the tendency to engage in a social risk (like trusting a stranger with personal information). Interestingly, the factor structure that best describes the covariance of ratings for behaviors does not perfectly match up with experimenter intuitions about the domains that the behaviors belong to. Thus, for example, in Study 1 of Weber et al. (2002), ethical risks like drunk driving and health risks like cigarette smoking loaded onto the same factor. Conversely, financial risks loaded

onto separate factors, one interpreted as investment risk and the other as gambling risk. Related work has found that although there are coherent "risk profiles" in the population, these do not cleanly subdivide based on the underlying domains examined by DOSPERT (Frey et al., 2023). For example, some individuals tend to be more likely to take both gambling and ethical risks, others are more likely to take gambling and investment risks, and yet others are more likely to take recreational and social risks. It is not clear why different items and individuals cluster together the way they do. Indeed, the psychometric paradigm, which infers underlying constructs (like domain structures and risk profiles) using only observed correlations across items or individuals, does not attempt to directly answer this question.

Researchers have also used DOSPERT and similar inventories to study how risk taking varies as a function of individual-specific variables, like age, gender, and personality (Blais & Weber, 2006; Dohmen et al., 2011; Frey et al., 2017, 2021; Harris & Jenkins, 2006; Highhouse et al., 2016; Josef et al., 2016; Weber et al., 2002). Unsurprisingly, people differ in the types of risks they take, and this can be predicted by their demographic and psychographic profiles. For example, Weber et al. (2002) and Harris and Jenkins (2006) found that men are more likely to engage in risky behavior than women in most domains, though this tendency disappears for social risks. In our own data, discussed below, we replicated this general tendency, though we also found important differences between behaviors within each domain. For example, men were less likely than women to engage in social risks like ending a friendship but more likely to engage in social risks like going door to door to sell a product. Again, this suggests that experimenter intuitions about the domain structure of risks may not be best suited for untangling the complex interplay between everyday behavior and individual-specific variables like gender.

Decision Modeling

An alternate approach to studying behavior is based on the decision modeling paradigm. While psychometric methods use insights from statistics for conceptualizing the structure of variance in a data set, decision models attempt to capture the processes that people use to make decisions using formal mathematical functions or computer algorithms (Bhatia et al., 2021; Busemeyer et al., 2019; He, Analytis, & Bhatia, 2022; He, Zhao, & Bhatia, 2022; see also Busemeyer & Diederich, 2010, for an accessible introduction to cognitive modeling). At the core of most decision models is the assumption that people represent everyday choice alternatives and behaviors in terms of their underlying attributes, features, and reasons. People also have preferences over these attributes, which take the form of attribute weights, and subsequently calculate the overall utilities of items through a weighted aggregation of their attribute values (Keeney & Raiffa, 1993). Formally, if we write the attribute values of a behavior i as a vector $\boldsymbol{x_i}$ and the attribute weights for individual j as a vector $\boldsymbol{w_j}$, this approach would describe the decision maker's utility (and subsequently propensity of engaging in the behavior) as $r_{ij} \sim \boldsymbol{w_j} \cdot \boldsymbol{x_i}$. Of course, people may not always use strictly linear models, as assumed here. For example, they may apply heuristics that simplify attribute weighting and aggregation, use nonlinear decision rules that interact the values of attributes, or aggregate attributes sequentially over time instead of all at once. Considerable research in psychology has focused on evaluating these diverse processes, as well as understanding how they interact with attention, memory, and other core properties of human cognition (Gigerenzer & Gaissmaier, 2011; Lee & Cummins, 2004; J. W. Payne et al., 1988; Roe et al., 2001; Weber et al., 2007; Zhao et al., 2022).

Regardless of the processes that people use to aggregate attributes, decision models provide a useful approach to quantitatively predicting and, more importantly, explaining both item- and individual-level differences in decision making. Since behaviors are described in terms of attributes, similarities and differences across these behaviors could be understood in terms of similarities and differences in attributes: Two behaviors that are highly correlated with each other likely evoke a similar set of reasons and have similar attribute profiles. Likewise, since preferences over behaviors are described in terms of attribute weights, similarities and differences across individuals can be understood in terms of similarities and differences in these weights: Two individuals that are highly correlated with each other likely care about the same reasons and place similar weights on the attributes as each other. To the extent that there is coherence in behavior within a domain (e.g., financial risk) or a social group (e.g., men) it would be because behaviors in that domain have similar attribute profiles and individuals in that social group have similar attribute weights.

Formally, we could quantify the similarity of two behaviors, i and i', as $\text{SIM}(\boldsymbol{x_i}, \boldsymbol{x_{i'}})$, and the similarity of two individuals, j and j', as $\text{SIM}(\boldsymbol{w_j}, \boldsymbol{w_{j'}})$, where SIM() is some similarity function that operates on pairs of attribute vectors (below, we will use cosine similarity, which measures similarities in terms of the directions of the vectors). If decision models are capable of describing item-level and individual-level differences in data, we would expect $\text{COR}(\boldsymbol{r_i^b}, \boldsymbol{r_{i'}^b}) \sim \text{SIM}(\boldsymbol{x_i}, \boldsymbol{x_{i'}})$ and $\text{COR}(\boldsymbol{r_j^p}, \boldsymbol{r_{j'}^p}) \sim \text{SIM}(\boldsymbol{w_j}, \boldsymbol{w_{j'}})$, respectively. In this way, the correlational structure observed in survey data (analyzed using standard psychometric methods) could be predicted using the attribute profiles of behaviors and preferences of individuals over these profiles (specified using decision models).

Such an approach could also explain why experimenter intuitions about domain structure may not fully predict observed correlations in behavior. Some behaviors could share attributes or reasons with behaviors across multiple domains, explaining why those behaviors do not cluster strongly with other behaviors in their own domain. If this is the case then we would also expect behaviors to covary systematically within domain, which should be predictable by the similarity of their attribute vectors. For the same reason, such an approach could explain why some demographic or psychographic variables are highly predictive of risk taking, why others are not, and why demographic and psychographic effects may hold for some domains but not others.

A final benefit of decision models is their ability to predict out-of-sample behavior. Since decision models describe an individual's preferences using decision weights, it is possible to infer these weights from ratings of one set of items, and apply them to a new set of items to describe how that individual will rate or evaluate the new items. This capability of decision models makes them especially valuable for applications to practical problems such as those involved in market research, managerial decision making, financial analysis, and public policy (see, e.g., Green & Srinivasan, 1990; Howard, 1988; Markowitz, 1959). For example, conjoint analysis measures the value consumers place on individual product attributes by presenting them with various combinations of these attributes, and then uses these insights to design new products that will maximize consumer preference and willingness to pay (Green & Srinivasan, 1990).

Integrating Psychometrics and Decision Modeling

Although decision models provide an elegant way of conceptualizing item- and individual-level differences in decision making, and potentially explaining nuances in empirical data, these models are not easy to apply to common real world behaviors, such as those measured using DOSPERT. Doing so is challenging as currently there is no way to specify the rich space of attributes and reasons that characterize common behaviors. What are the $\boldsymbol{x}_i$s that describe everyday risks like investing in a speculative stock, mountain climbing, eating high cholesterol foods, shoplifting, and trusting a stranger with personal information? Without solving this practical problem, decision models remain largely restricted to artificial laboratory settings and highly stylized theoretical applications.

It is worth noting that many applications of DOSPERT-type inventories do ask people to rate their perceptions of the expected costs and benefits of various behaviors, and use these ratings (often in linear models) to predict ratings of behavioral propensities and differences across participants (see, e.g., Weber et al., 2002 for an example). Although such applications can be considered simple decision models, they are unable to predict the reasons that cause people to perceive some behaviors as having high or low costs and benefits, and why different people may vary in these perceptions. For this reason, such approaches are inherently limited for out-of-sample prediction, as projecting unseen items onto attributes like costs and benefits requires obtaining participant ratings on those attributes.

It is also possible to think of the factor structure revealed by psychometric analysis as quantifying the attributes at play in decision models. In the case of DOSPERT this would involve representing each risky behavior in terms of the five of six dimensions that capture the variance in ratings data. This is, in fact, the dominant approach in the closely related problem of risk perception. In pioneering work, Fischhoff et al. (1978) elicited human ratings for a set of nine experimenter-generated reasons and used these ratings to explore what causes people to see an activity or technology as risky (see also Slovic, 1987). They found that people's ratings clustered into two factors, one capturing the familiarity of risks and the knowability of their consequences, and the other capturing the degree of dread and the potential for fatality of their consequences.

Although risk perception is closely related to risky decision making, it may not be possible to use the psychometric approach for the latter task as risky decisions involve a complex set of social, legal, financial, moral, and emotional variables, that would be hard to capture using a small set of experimenter-generated reasons. As an example, consider the following two DOSPERT items: leaving a child alone at home and pirating software. Although both have legal repercussions, the former also includes concerns about the child's safety, the desire to make the child independent, and the parents' need to work and do errands, whereas the latter includes concerns about software quality and security, factors involving convenience and affordability, as well as a desire to support open source products. Each of these reasons reflects several nuanced considerations, involving both the decision maker and other stakeholders in the decision, and many of these considerations are highly specific to the behavior itself. Although some of these reasons could be generated by a theorist or experimenter (as Fischhoff et al. did for risk perception), it would be impossible to specify the thousands of possible reasons that could play a role in any possible behavior, and to moreover to get human participants to rate these reasons to derive quantitative representations for the behaviors with psychometric techniques. Additionally, as discussed above, any psychometric method for measuring attribute structure would be fundamentally handicapped in making out-of-sample predictions, as it relies on the ratings of the behaviors themselves to specify their underlying attributes.

One promising new approach to examining the reasons underlying everyday risk uses open-ended natural language listings of reasons. For example, Steiner et al. (2021) examined mental representations of risk using a thought generation method in which participants were prompted to list all reasons crossing their minds while evaluating themselves on a general risk taking measure. Similarly, Arslan et al. (2020) asked participants to list specific events, behaviors, or situations they considered when rating their own risk-taking tendencies. These responses were then manually coded and quantified by either the participants themselves (Steiner et al.) or separate coders (Arslan et al.). Both articles found that the coded self-reports predicted risk taking propensity and provided insight into core dimensions influencing people's assessments of their own riskiness.

While insightful, these two approaches do not fully solve the key challenge of building predictive decision models for established psychometric inventories like DOSPERT, which is the primary goal of this article. First, they examine only general risk-taking tendencies and do not explain variability across specific items and domains of risky decision making. Second, they rely on participants to generate reasons themselves and on manual human coding of listed reasons, which precludes out-of-sample prediction. Thirdly, human coding yields relatively impoverished representations unable to capture rich diversity in participant-generated reasons, and thus does enable fully fledged quantitative modeling. Overall, the techniques in Steiner et al. and Arslan et al. correlate self-reports with behavior, which facilitates high-level psychological interpretation, but does not provide us with a complete computational framework for predicting specific risky decisions taken from different domains.

Nonetheless, these self-report approaches represent an important first step by demonstrating the utility of open-ended natural language rationales and reasons for understanding risk taking. If aspects could be automatically generated for wider item arrays like DOSPERT, and their nuances quantitatively encoded without extensive human effort, then it could be possible to build types of predictive models necessary for capturing the interitem and interindividual correlations in risk taking currently analyzed only using psychometric techniques.

LLMs

This final barrier can be overcome with LLMs. There has been a growth of digitized linguistic data over the past few years, as human discourse increasingly takes place on the internet. Researchers have begun training machine learning models on this data to derive representations for words and sentences. These methods all work using the following intuition: Words and sentences that appear in similar contexts have similar meanings and can thus be given similar representations. The representations take the form of neural network connection weights, which are multidimensional vectors (often known as embeddings). In this way, nearly any concept or construct can be vectorized as long as it is describable using words and sentences (Brown et al., 2020; Devlin et al., 2018; Mikolov et al., 2013; Reimers & Gurevych, 2019)

Building on this approach, researchers have begun using LLMs to specify representations in psychological applications. These applications

include models of similarity judgment (e.g., Landauer & Dumais, 1997; Mandera et al., 2017), memory retrieval (Hills et al., 2012; Richie et al., 2023), implicit bias (Bhatia & Walasek, 2023; Caliskan et al., 2017), semantic cognition (Bhatia & Richie, 2023; Lu et al., 2019), and, most relevant to this article, models of human judgment and decision making (see, e.g., Bhatia & Aka, 2022; Bhatia et al., 2019 for reviews). For example, in Bhatia (2019), Gandhi et al. (2022), and Bhatia et al. (2022), we have used vector representations for words and short phrases to predict perceptions of the riskiness of technologies and activities, healthiness of foods, and leadership qualities of individuals (see also Richie et al., 2019). Although the judgment targets in these articles are simple one or two-word objects (like nuclear power, orange juice, and Nelson Mandela, respectively) and these articles have not attempted to systematically test the models on their ability to describe item- or individual-level differences, these successful applications nonetheless illustrate the promise of LLM-derived representations for judgment prediction. In other relevant work, Zhao et al. (2022) have used sentence vector representations to model the memory processes at play in the retrieval of reasons in a small set of everyday decisions. Relatedly, Aka and Bhatia (2022) have used sentence vector representations to predict how people evaluate descriptions of common diseases and health states, and Singh et al. (2022) have used sentence vector representations for common verb phrases to predict people's self-reported propensity to engage in the behavior corresponding to the phrases. Finally, Abdurahman et al. (2024) have shown that sentence vector representations for personality items can predict people's ratings of those items. Although the precise models, data sets, and research questions in the above articles are very different to those in the current article, they nonetheless show the applicability of sentence vector models for quantifying the complex attributes and considerations at play during naturalistic deliberation.

All of the methods discussed in this section use vector representations of words and sentences obtained from connection weights in deep neural networks trained to generate high-quality word and sentence representations. This is not the only way to use LLMs to extract attribute representations for objects, concepts, and behaviors. A complementary approach is to explicitly ask generative language models to list reasons for or against engaging in a given behavior. Generative language models, like generative pretrained transformer (GPT) (Brown et al., 2020), are a special type of LLM that have been optimized for producing language. Although these models also represent words and sentences as vectors in deep neural network layers, their ability to generate linguistic output can be used to give these high-dimensional representations human-interpretable linguistic labels. Building off the pioneering work of Arslan et al. (2020), Steiner et al. (2021), and others, we will use generative LLMs to automatically extract reasons for and against several real-world behaviors, which we will transform into multidimensional vector representations amenable to decision modeling. We will then use these to both predict and interpret the main dimensions of behavioral variability in our data sets.

Computational Method

Algorithmically Derived Reasons

The behaviors used in our experiments were taken from the DOSPERT inventory or an extended version of this inventory (described in detail below). All items were in the form of short natural language phrases. We used four methods to build quantitative multi-attribute representations for these phrases. The first two of these methods relied only on the behavior phrase (e.g., investing in a speculative stock), whereas the second two of these methods relied on machine-generated reasons for and against each behavior (e.g., speculative stocks offer the potential for greater returns compared to low-risk investments and speculative stocks carry a greater risk than most other investments, due to their unpredictable nature and potential for rapid declines in value). We generated these reasons by querying GPT-3.5 (text-davinci-003) with the following prompt: "What are reasons for or against [BEHAVIOR PHRASE]? List five reasons for each, and number them." We used the OpenAI API instead of ChatGPT since the API made it possible to algorithmically run and record the responses for a large set of queries. The API also has fewer guardrails, and is willing to give controversial responses, like reasons in favor of drunk driving or consuming heroin (ChatGPT, by contrast, refused to generate potentially unethical advice).

Overall, we were quite surprised by the quality of the reasons generated by GPT, which was able to give clear and comprehensive reasons for all behaviors and was even able to generate insightful reasons that we had not thought of. We used GPT-generated reasons to predict behaviors in Study 1. In Study 2, we directly compared GPT's reasons with those generated by humans and also used these reasons to interpret the main sources of item- and individual-level variability. GPT-generated reasons for the items are available in the Open Science Framework repository for this project (https://osf.io/3y6ku/).

Vectorizing Behaviors

We also considered two different techniques for quantifying the texts (the language used to describe the behavior or the GPT-generated reasons for and against the behavior) discussed in the prior section. The first of these was a bag-of-words Word2Vec model. In this model, the text was first preprocessed by tokenizing it into individual words and removing any stop words and punctuation. In this way the text was simplified into only the set ("bag") of its component words. Then, each word was assigned a 300-dimensional vector representation obtained using the popular Word2Vec model (Mikolov et al., 2013). Word2Vec is a neural network that uses the co-occurrence statistics of words in large language corpora (books and news articles) to derive representations that capture semantic relationships between words. Similar words are given similar representations. Although Word2Vec's representations are only for words, the bag-of-words technique generates a vector representation, $\boldsymbol{x_i}$, for behavior i, by averaging the individual word vectors in the behavior phrase or the set of GPT-generated reasons for that behavior.

Our second technique for vectorizing behavior relied on the sentence-bidirectional encoder representations from transformers (SBERT) model (Reimers & Gurevych, 2019). Unlike the prior method, which treats sentences as bags of words, SBERT considers the contextual and syntactically constrained meaning of words within a sentence. It utilizes a neural network architecture that fine-tunes pretrained transformer models like BERT on a sentence similarity task. By training on large amounts of data, SBERT learns to encode sentences into vectors that capture their meaning. These vectors enable efficient computation of sentence similarities and have been shown to be successful for various downstream natural language processing tasks. The SBERT model used in the current

article was based on the RoBERTa model (Devlin et al., 2018; Liu et al., 2019), which was trained on a large corpora of online text data. We queried this model using huggingface's API (sentence-transformers/all-roberta-large-v1). Recall that we applied this model both to textual descriptions of the behavior and to GPT-generated reasons for and against the behavior. In the former case, we just passed the item's text (e.g., investing in a speculative stock) through the SBERT model. In the latter case, we first parsed GPT's textual output into a list of 10 reasons (five for and five against). Each of these reasons (e.g., speculative stocks carry a greater risk than most other investments …) was passed through SBERT and the vectors for each of the 10 reasons were averaged. Both approaches generated 1,024 dimensional representations for each behavior, which we write as $\boldsymbol{x_i}$ for behavior i.

Overall, the set of models considered in this article has a 2×2 factorial structure, which varies the linguistic depiction of the behavior (the item's behavior phrase or GPT-generated reasons for the behavior) and the technique for vectorizing this linguistic depiction (bag-of-words Word2Vec or SBERT). We thus refer to these four models as Word2Vec-Items, Word2Vec-Reasons, SBERT-Items, and SBERT-Reasons, respectively. Figure 1A summarizes the four LLM methods used in our analysis pipeline.

It is worth noting that using LLMs to generate reasons for behaviors and then quantifying these behaviors with other LLMs, as in the SBERT-Reasons model, has, from a statistical perspective, a certain amount of redundancy, since it is merely doing a transformation in LLM's vector space without adding any additional information. Thus, we would not expect SBERT-Reasons and Word2Vec-Reasons to have higher predictive accuracy than SBERT-Items and Word2Vec-Items, respectively. Despite this, we included all four models in our analysis as GPT-generated reasons are useful for interpreting the sources of variance modeled using our framework and additionally allow us to understand the processes that individuals go through as they deliberate.

Fitting Decision Models

Our goal was to use the vectors $\boldsymbol{x_i}$, obtained from the four methods described in the previous section, as inputs into decision models. Although the dimensions of these vectors may not have direct interpretability, similar behaviors nonetheless possess similar vectors and have similar values on each of the (300 for Word2Vec or 1,024 for SBERT) dimensions. Additionally, people's preferences for these behaviors can be understood in terms of the weights, $\boldsymbol{w_j}$, that they assign to the dimensions. Consequently, a decision model can be constructed to predict the rating for behavior i for individual j, using the following equation: $r_{ij} \sim \boldsymbol{w_j} \cdot \boldsymbol{x_i}$. In this equation, the weights correspond to the preferences individuals have for each dimension, while the dimensions represent the characteristics or aspects of the behavior captured by the vectors. By assigning appropriate weights to the dimensions, the decision model can effectively predict or evaluate behaviors based on the given vector representations.

To determine each individual's weights in the decision model, we employed a ridge regression modeling technique. This is a regularized regression method that aims to minimize the sum of squared errors in a linear model while incorporating a penalty term. This penalty term restricts the coefficients of the weights, promoting a balance between accuracy and simplicity in the model. Ridge regression helps mitigate multicollinearity and overfitting, which are potential issues in settings with high-dimensional predictors (as in our article). For this reason, it has performed particularly well in applications that use text vectors to predict human responses (e.g., Richie et al., 2019). We implemented the ridge regression using the sci-kit learn machine learning library (Pedregosa et al., 2011), with regularization penalties, a, in the set [0.001, 0.01, 0.1, 1, 10].

We evaluated our models using leave-one-out cross-validation (LOOCV). LOOCV measures the performance of a model by leaving out one data point from the training set and using it as the test set. The process is repeated for each data point, ensuring that every instance is used as both training and test data. LOOCV provides an unbiased estimate of the model's performance as it evaluates how well the model generalizes to unseen data. We applied LOOCV to each individual separately. Thus, for each individual, we fit a decision model to all except for one of their ratings and evaluated the model on the held-out rating. We repeated this process for each rating given by each individual. This technique is illustrated in Figure 1B.

As part of our tests, we calculated how well similarities between $\boldsymbol{x_i}$s or between $\boldsymbol{w_j}$s predicted correlations between behaviors and between individuals, respectively. We used cosine similarity to measure vector similarity. This is defined as the cosine of the angle of two vectors. Specifically, according to cosine similarity, $\mathrm{SIM}(\boldsymbol{a},\boldsymbol{b}) = (\boldsymbol{a}\cdot\boldsymbol{b})/(\|\boldsymbol{a}\|\cdot\|\boldsymbol{b}\|)$. Cosine similarity is equal to +1 for two vectors that are in the same direction, −1 for two vectors that are in the opposite direction, and 0 for two vectors that are orthogonal to each other.

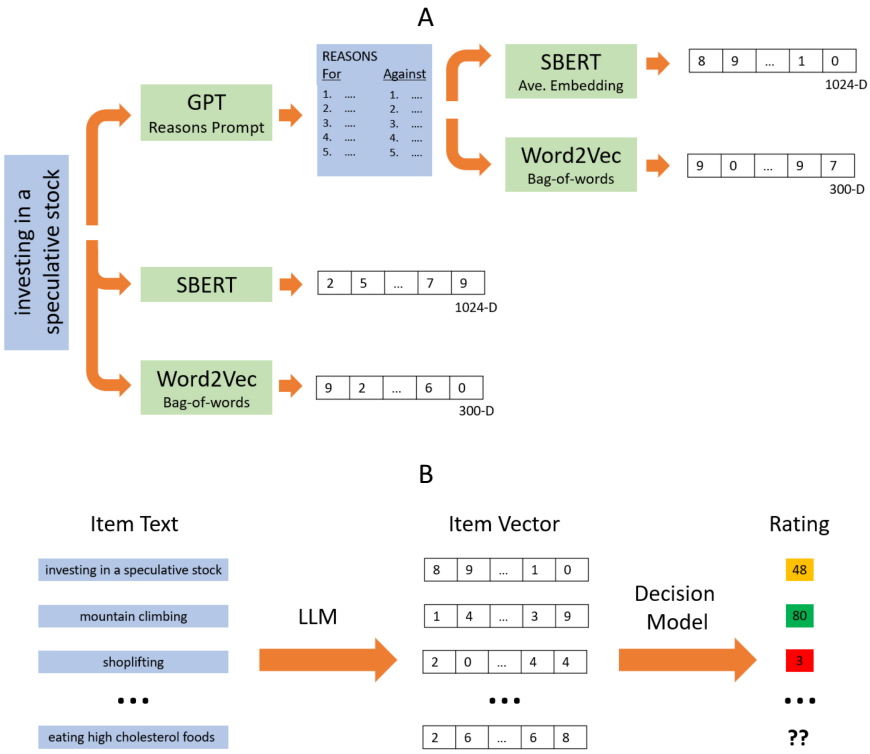
Additional Tests

To better understand the properties of our modeling approach we also considered seven additional model-based tests. The first of these used a random vector model. This model replaced the vectors $\boldsymbol{x_i}$ in the above pipeline with randomly generated numbers, each sampled from a standard normal distribution. The dimensionality of these random vectors was the same as SBERT's (1024) dimensionality, allowing us to test whether our model's performance is simply due to the flexibility inherent in high-dimensional predictors.

The second test used a DOSPERT labels model, which replaced the vectors $\boldsymbol{x_i}$ in the above pipeline with five-dimensional binary vectors indicating the experimenter specified domain classification for the items. Thus, for example, an item that was considered (according to the experimenter) to be in the financial domain was given a vector [1, 0, 0, 0, 0] whereas a model that was considered to be in the recreational domain was given a vector [0, 1, 0, 0, 0]. Although this model can be seen as a weighted additive multiattribute decision model, with best-fitting weights corresponding to the participant's weighting of the five DOSPERT domains, it also has an alternative, simpler interpretation: At its core, this model simply predicts the rating of an out-of-sample target item by averaging the participant's ratings for items in the training data that share the target item's domain. Thus, an out-of-sample item from the financial domain will be given roughly the average rating of financial items that the model was trained on. Overall, the DOSPERT labels model serves as a benchmark for our approach as it evaluates the predictive power inherent in the intuitively derived taxonomy of the DOSPERT inventory.

While the previous two tests replace the predictors used in the analysis pipeline, the third test replaces the training items. In

Figure 1
Overview of Approach



Note. (A) The four LLM methods for vectorizing behaviors. We used SBERT and bag-of-words Word2Vec to vectorize text and applied these either directly to the behavior text or to GPT-generated reasons for and against engaging in the behavior. (B) Description of an LLM-based decision model for making out-of-sample rating predictions. Here item vectors, obtained from one of the four LLM methods, are passed through an individual-specific decision model, in which attribute dimensions are weighted and aggregated to predict ratings. The model's weights are fit on the participant's ratings for all items except for one test item, and the model's accuracy is tested on the held-out rating for the test item. GPT = generative pretrained transformer; SBERT = sentence bidirectional encoder representations from transformers; Ave. = average; LLM = large language model. See the online article for the color version of this figure.

particular, unlike our main analysis which uses leave-one-out cross validation on our full data set of items, the original DOSPERT items test trains our model only on the ratings for original DOSPERT items and then tests it on ratings for extended DOSPERT items (see details of the DOSPERT extension below). This analysis helps us evaluate the sensitivity of our model performance to data set size and content (the original DOSPERT inventory has far fewer, and somewhat dated, items than our extended inventory). Our next three tests are variants of the DOSPERT items test that randomly sample either 25%, 50%, or 75% of the items in our extended DOSPERT inventory for training purposes. As with the DOSPERT items test, fitting our model on subsets of the extended DOSPERT inventory and testing it on the held-out items helps us evaluate the dependence of our modeling pipeline on data set size.

Our final test contrasts the above analyses with a psychometric model that is trained on the full data set. To implement this DOSPERT in-sample model, we performed a factor analysis on the ratings of all participants on all items, in order to represent each item as a vector of five latent factors. This list of five latent factors was used to specify the $\boldsymbol{x_i}$ in the above pipeline, so that, for a given participant, we first regressed the ratings on the five latent factors, and then used the best fit model to predict the ratings for a test item. Since the $\boldsymbol{x_i}$s in this analysis (i.e., the vector representations for the items) are derived from the participant's ratings for those items, this is not an out-of-sample model. Nonetheless, this model is valuable as a benchmark since it represents the maximum predictive power that can be obtained using psychometric techniques. In this sense, it serves as an upper bound on predictive accuracy in our tests.

Interpreting Reasons

The above pipeline is formulated for out-of-sample prediction but does not help us understand the higher-level constructs that drive variability in ratings for items across individuals. Thus, in addition to using the above LLM representations of reasons to predict ratings, we also used LLM representations to code the items on 18 different interpretable themes. We generated these themes separately for each

of the five DOSPERT domains. Overall, we came up with between three and five themes that characterized the different types of items and attributes in each domain. Then for each of the themes we generated 10 different associated words in order to specify a lexicon or dictionary for the theme (e.g., Tausczik & Pennebaker, 2010) (see Table 1). Subsequently, we obtained the Word2Vec vectors for each of the words in the theme's lexicon and averaged these vectors to obtain a single vector representation for the theme. Finally, we measured the extent to which each theme occurred in the reasons for each item by calculating the cosine similarity of the theme vectors to the Word2Vec-Reasons vector representations for the items (these, as discussed above, are the average of the Word2Vec vectors for the individual words in the reasons for the items). Recall that we write the vector for item i as $\boldsymbol{x_i}$. Thus the degree to which theme k is present in item i is simply $\text{SIM}(\boldsymbol{x_i},\boldsymbol{t_k}) = (\boldsymbol{x_i}\cdot\boldsymbol{t_k})/(\|\boldsymbol{x_i}\|\cdot\|\boldsymbol{t_k}\|)$, where $\boldsymbol{t_k}$ is the theme vector for a theme k. Intuitively, this similarity measure captures the extent to which the words used in a given theme's lexicon are semantically similar to the words in a given item's GPT-generated reasons. Our approach is based on the distributed dictionary method, which has been shown to be particularly useful for coding small amounts of text on complex psychological properties (see Garten et al., 2018 for details).

Comparing Machine- and Human-Generated Reasons

Lastly, we attempted to compare GPT-generated and human-generated reasons, in order to evaluate whether our approach can predict (a) the different reasons that people generate for different items, and (b) the different reasons that different people generate for the same item. The first of these analyses involved simply comparing the similarity of GPT-generated and human-generated reasons for items. Here we vectorized GPT-generated and human-generated reasons using SBERT and tested whether the vector of the GPT-generated reasons for an item had a higher cosine similarity with the vectors for participant-generated reasons for that item, relative to participant-generated reasons for other items. The second analysis was somewhat more complex as it needed to consider the weights different participants place on different dimensions, that is, participants' idiosyncratic preferences. Thus, this analysis used a modification of the above approach, which examined the cosine similarity between the weighted attribute vectors for a given participant–item combination and the vector for the reasons that the participant generated for the item.

Table 1
List of Themes, and Examples of Words, Used in Study 2 Interpretative Analysis

Domain	Themes	Words
Ethics	Law	Arrest, illegal, prison …
	Norms	Obligations, expectations, customs …
	Morality	Ethics, integrity, virtue, …
Health	Mental health	Well-being, anxiety, depression …
	Injury	Fracture, wound, sprain …
	Disease	Infection, virus, bacteria …
	Nutrition	Diet, vitamins, protein …
Social	Family	Parents, children, siblings …
	Friendship	Friend, acquaintance, solidarity …
	Romance	Love, sex, marriage …
	Career	Profession, employment, ambition …
	Education	Learning, school, knowledge …
Recreation	Adventure	Journey, thrill, expedition …
	Pleasure	Enjoyment, fun, luxury …
	Relaxation	Rest, leisure, comfort …
Financial	Investment	Stocks, bonds, portfolio …
	Personal finance	Budgeting, savings, expenses …
	Gambling	Betting, casino, odds …

Study 1

The goal of Study 1 was to evaluate the utility of LLM-based decision models for capturing item-level and individual-level variability in risky decision making. For this reason, it administered an expanded version of the DOSPERT scale to a sample of U.S. participants, and attempted to fit individual-level decision models to the observed data. We evaluated the accuracy of various LLM-derived behavior representations in predicting out-of-sample participant responses. Importantly, we also tested how well similarity in LLM representations predicted observed correlations across items and domains, as well as how well similarity in best fit attribute weights predicted correlations across individuals and groups.

Method

Participants

We recruited 150 participants (79 female-identifying, 67 male-identifying, four other) from Prolific Academic, an online platform for conducting research with human subjects. The participants were aged between 18 and 72 years ($M = 37.64$, $SD = 13.69$), and were U.S. citizens fluent in English. All participants gave informed consent before taking part in the study. The study was approved by the UPenn Institutional Review Board (823184). We selected a sample size of 150 since it is a nice round number in line with research practices in the field.

Procedure

The participants were asked to rate their relative likelihood of engaging in 150 behaviors on a scale of 0 (*much less likely than others*) to 100 (*much more likely than others*). These behaviors were taken from the extended DOSPERT inventory, detailed below. The 150 behaviors were presented in a random order on the same screen, and the participants used a slider to indicate their ratings. The slider had a default start position of 50, which meant that the participants were equally likely as others to engage in the behavior. The participants could change their ratings as many times as they wanted before submitting their responses. Embedded in the set of items was an attention check item which asked participants to respond with a rating of 26. The five participants who failed the attention check were eliminated from the study, resulting in a final sample size of 145. Note that we had forgotten to preregister the attention check and repeating our analysis without this exclusion criteria results in no change to the results.

After rating all the behaviors, the participants completed a demographic questionnaire that asked about their age, gender, education level, and occupation. They also completed a Big Five personality inventory that measured their levels of extraversion, agreeableness, conscientiousness, neuroticism, and openness to experience. The inventory consisted of 10 items that were rated on a 5-point Likert scale from 1 (*strongly disagree*) to 5 (*strongly agree*) (Rammstedt & John, 2007).

Materials

The 150 items used in our study were taken from the extended DOSPERT inventory. As with the original DOSPERT inventory, the extended DOSPERT inventory covered five established domains of risky choice: financial, ethical, social, health, and recreational. The pool of behaviors in this inventory was generated by merging and expanding the original DOSPERT-30 and DOSPERT-40 inventories (Blais & Weber, 2006; Weber et al., 2002), so that each of the five domains of DOSPERT had a total of 30 behaviors each. Out of the 150 items in our extended inventory, 44 were taken from DOSPERT-30 and DOSPERT-40, whereas the remaining were new items.

While extending the original DOSPERT inventory, we took into consideration several factors. First, we included several contemporary risks that were not present in the original DOSPERT inventory (e.g., investing in cryptocurrency). Secondly, we attempted to add items that maximized the diversity of the types of behaviors in each domain. For this reason, we did not closely replicate existing items, rather we tried to include items and themes that were excluded from the original inventory (examples of new and distinct items are speaking up against injustice at your workplace, and spending a large amount of time playing video games). Finally, we removed nuanced quantitative details from DOSPERT items (e.g., investing 5% of your annual income in a very speculative stock was replaced with investing in a speculative stock) and removed items that were very similar to and considered safer alternatives to other items in the inventory (e.g., instead of including both investing in a speculative stock and investing in a moderate growth diversified fund, which were both present in the original DOSPERT inventory, our extended inventory used only the former item). We preregistered the behaviors before launching the study, and the full set of behaviors is provided in the OSF repository for this project.

Transparency and Openness

We have reported how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. The analysis was performed in Python. The study design was preregistered at https://osf.io/amves/. Code and data are available at https://osf.io/3y6ku/.

Results

Predictive Accuracy

First, we wanted to assess the effectiveness of our approach in predicting participant ratings. Recall that we used four different models for obtaining representations. These models had a 2×2 factor structure based on whether they used the items' behavior phrases or GPT-generated reasons for the items, as well as whether they used the bag-of-words Word2Vec model or the SBERT model for quantifying the texts. Each model's representations were passed through a ridge regression, which allowed for one of five flexible regularization penalties. To evaluate the models, we employed LOOCV on a participant level. This involved calculating the model's out-of-sample prediction for each rating obtained from each participant and comparing it with the observed data.

Figure 2 illustrates the accuracy rates of these predictions. In Figure 2A, we present the average correlation between predicted and observed ratings across all participants. Figure 2B presents the average mean square error (*MSE*) between each model's out-of-sample predictions and the observed ratings. Both figures show that SBERT-Items was the best performing model, and that this model performed the best with a ridge penalty value of $a = 1$ (the default parameter in the scikit-learn library). Figure 2C shows the distribution of participant-level out-of-sample correlations for this model. Here we can see that SBERT-Items achieved positive correlations for all except for one participant in the data set. The average correlation across participants is .43, which is statistically greater than 0, $t(144) = 43.97$, $p < .001$, 95% confidence interval (CI) = [0.41, 0.45].

Additional Tests

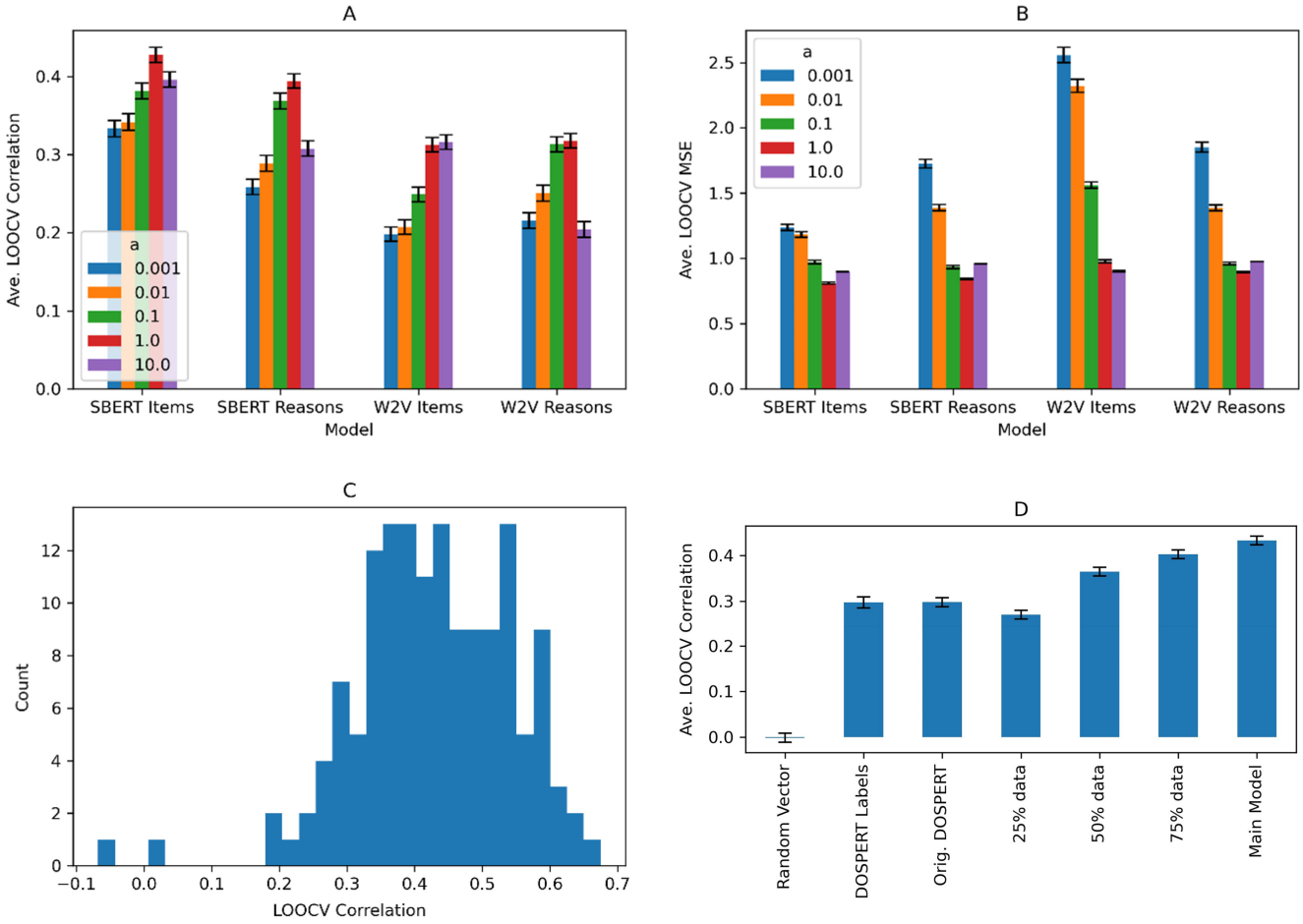
We also conducted a number of additional tests to evaluate the predictive accuracy rates of our model. The results of these tests are summarized in Figure 2D. Our first test used a 1,024-dimensional random vector model which randomly sampled attribute representations from a standard normal distribution. We found that this model performed very poorly, achieving statistically zero correlation in out-of-sample tests, average r across participants $= -0.001$, $t(144) = -0.16$, $p = .87$, 95% CI = [−0.02, 0.01]. This shows that it is not the dimensionality of the LLM vectors but their resemblance to human semantic representations that matters during prediction.

Our second test used DOSPERT labels to generate binary attribute vectors that described the domains the items belonged to. Intuitively this model predicts a participant's rating of an out-of-sample item using their ratings on other items in the same (experimenter-specified) DOSPERT domain. We trained this model on each participant's responses with a standard linear regression and tested it on out-of-sample data using LOOCV. This model performed decently, achieving an out-of-sample correlation of $r = .30$. Although this is significantly different to zero, $t(144) = 24.32$, $p < .001$, 95% CI = [0.27, 0.32], it is still considerably worse than our main model, showing that there is meaningful psychological nuance within each domain that LLM representations (but not experimenter intuition) can capture.

We also assessed the predictive ability of the original DOSPERT items for the extended DOSPERT items. To do this, we replicated the above analysis with a slight modification: We trained the best performing model, SBERT-Items (with ridge penalty $a = 1$), on each participant's responses to the 44 existing DOSPERT items in our scale. We then calculated its predictive accuracy, measured in terms of out-of-sample correlation, on the 106 new items in our scale. We found that the average correlation between the predicted responses and the observed responses to the new items was .30. This is a significant deviation from zero, $t(144) = 30.06$, $p < .001$, 95% CI = [0.28, 0.32], and shows that our method applied to the original DOSPERT scale can accurately predict responses to new behaviors.

The next three tests used a variant of the original DOSPERT items test. In particular, these three tests randomly sampled either 25%, 50%, or 75% of the extended DOSPERT inventory and used a participant's ratings of these items to predict their rating of a held-out item with the SBERT-Items model (with ridge penalty $a = 1$). We found that the average correlation between the predicted responses and the observed responses was .27, $t(144) = 28.10$, $p < .001$, 95% CI = [0.25, 0.29]; .36, $t(144) = 36.17$, $p < .001$, 95% CI = [0.34, 0.38]; and .40, $t(144) = 39.82$, $p < .001$, 95% CI = [0.38, 0.42], for the three models, respectively. These successful predictions

Figure 2
Predictive Accuracy



Note. (A) Average LOOCV correlation of four models across participants, with varying ridge regularization penalties *a*. (B) Average LOOCV *MSE*s of models. (C) Histogram of participant-level LOOCV correlations for the best performing model, SBERT-items. (D) Average LOOCV correlation obtained from six alternate tests along with that for the best performing model from the main analysis. Note that we do not display the results of the DOSPERT in-sample model in Figure 2D as its predictions are not out-of-sample (this model achieved a correlation of .63). Error bars indicate ± 1 *SE* in Panels A, B, and D. Ave. = average; LOOCV = leave-one-out cross-validation; W2V = Word2Vec; DOSPERT = Domain-Specific Risk Taking scale; Orig. = Original; *MSE* = mean square error. See the online article for the color version of this figure.

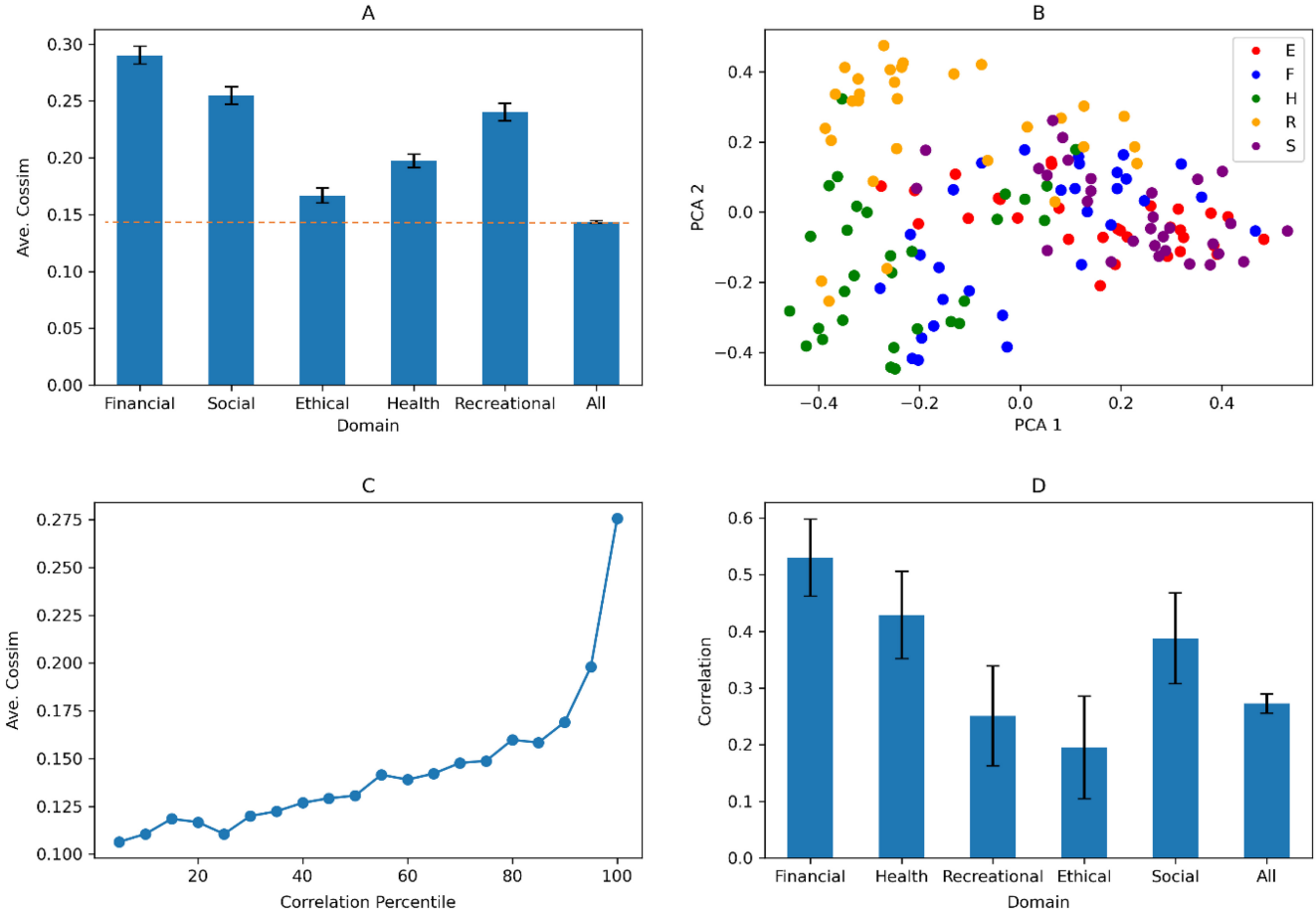
show that our approach is able to achieve good performance even when using drastically less data. For example, the model that used only 25% of the data did only slightly worse than the DOSPERT labels model in the second test, demonstrating that slightly more than a quarter of the data is needed to replicate prior approaches that rely on experimenter intuition. This analysis should address any reservations about undue dependence on data set size.

Our final test attempted to contrast the above results with the DOSPERT in-sample model that used ratings from all participants on all items to derive latent factor representations for the items. As discussed above, this model does not attempt out-of-sample prediction. Nonetheless, it is useful as a benchmark, as it quantifies the upper bound on predictive accuracy that can be obtained from psychometric techniques applied to the full data set. We found that this model achieved an average correlation between the predicted responses and the observed responses of .63, $t(144) = 6.84 \times 10^{16}$, $p < .001$, 95% CI = [.063, 0.63].

Item Variability

One of the key reasons for the high accuracy rates observed in our study is the rich representational structure embedded within our vectors. These vectors have the ability to capture item similarity, meaning that behaviors that are given similar ratings by participants are assigned similar attribute vectors by our LLMs. To formally test this explanation, we measured the average cosine similarity between pairs of item vectors within each domain for our best performing model, SBERT-Items. We also measured the average cosine similarity between all items in our inventory. The results of this analysis are in Figure 3A, which shows that the pairwise cosine similarity between items within a domain is significantly higher than the pairwise cosine similarity between items across the full inventory. In other words, the SBERT-Items model is able to capture the domain structure of the DOSPERT inventory by projecting items from each domain onto separate regions of its representational space.

Figure 3
Item Variability



Note. (A) Average pairwise item cosine similarity within the five DOSPERT domains as well as across all items in the inventory (which is also indicated with the dashed line). (B) Scatter plot of first and second principle components of item vectors, colored by domain. (C) Average pairwise cosine similarity of item pairs as a function of their pairwise correlation in participant ratings. Items are binned into 20 equally sized groups based on the participant correlations. (D) Correlations between pairwise item cosine similarity and pairwise item participant correlations within each of the five domains as well across the full data set. The results in this figure use vector representations from the SBERT-Items model. Error bars indicate $\pm$ 1 *SE* in Panel A and 95% CIs in Panel D. Ave. Cossim = avearge cosine similarity; PCA = principle component; E = ethical; F = financial; H = health; R = recreational; S = social; DOSPERT = Domain-Specific Risk Taking scale; SBERT = sentence bidirectional encoder representations from transformers; CIs = confidence intervals. See the online article for the color version of this figure.

To provide a visual depiction of these patterns, we generated a scatter plot using the first two principal components dimensions of the SBERT-Items model's vectors. This position of items on these two components is shown in Figure 3B. This plot shows domain coherence, with items from each domain having a high degree of proximity to other items in their respective domains. The scatter plot also reveals a graded structure, indicating continuity in the underlying representations. For example, the original DOSPERT's health risks, not wearing a seatbelt and walking alone at night in an unsafe area of town are close to recreational and social risks, respectively, whereas other health risks from the original DOSPERT inventory, like regularly eating high cholesterol foods, are not. This graded structure indicates that our vector representations may capture nuanced variations in similarity that go beyond discrete domain categories.

To further assess our item vectors' representational quality, we compared the pairwise cosine similarity between the 11,175 unique item pairs in the extended DOSPERT inventory to the human correlations observed in the ratings of those items, for our best performing model. Formally, we tested whether $\mathrm{COR}(\boldsymbol{r}_i^b, \boldsymbol{r}_{i'}^b) \sim \mathrm{SIM}(\boldsymbol{x}_i, \boldsymbol{x}_{i'})$, where $\boldsymbol{x}_i$ is item is 1,024-dimensional SBERT-Items vector, $\boldsymbol{r}_i^b$ is the 145-dimensional vector of ratings given to item i by our 145 participants, and i and i' are indices in range (1, 150) with $i \neq i'$. We found that the correlation between $\mathrm{SIM}(\boldsymbol{x}_i, \boldsymbol{x}_{i'})$ and $\mathrm{COR}(\boldsymbol{r}_i^b, \boldsymbol{r}_{i'}^b)$ across all pairs is .27, which is significantly different to zero ($p < .001$, 95% CI = [0.25, 0.28]). This is illustrated in Figure 3C. For visual clarity, we divided the 11,175 item pairs into 20 equally sized bins, categorized based on the percentile ranking of their human correlation values. The figure displays the average cosine similarity of items pairs within each bin. The findings depicted in Figure 3C reveal a positive relationship between the human correlations between items and the cosine similarities of items. Bins that contain item pairs with higher human correlations also exhibit higher cosine similarities between

those items' vectors. Additionally, this relationship has a convex trend, indicating that the bins with the highest human correlation values have proportionally higher vector similarities.

Finally, to assess whether our item vectors effectively capture human ratings beyond discrete experimenter-generated domains, we repeated the analysis depicted in Figure 3C within each of the five domains. For instance, we calculated the pairwise correlations among all recreational risks and compared these with the cosine similarities of the SBERT-Items vectors for recreational risks. If human behavior were solely predicted by domain overlap, we would expect zero correlations. This is because there would be no systematic correlational structure within each domain to explain. However, if there is graded structure that extends beyond discrete domains, and if our vectors successfully capture this structure, we would expect to observe positive correlations. Figure 3D shows that we obtained significantly positive correlations for all domains. For example, drinking sugary drinks and eating processed foods received similar ratings across individuals (rating correlation of .60) and were also given similar vectors by our model (cosine similarity of .67). By contrast, not brushing regularly and using a ultraviolet tanning machine received dissimilar ratings across individuals (ratings correlation of .02) and also were given dissimilar vectors by our model (cosine similarity of .05). This is despite the fact that these four behaviors are all taken from the health risk domain. Overall, these findings provide strong evidence that there are underlying structures that extend beyond the boundaries of discrete domains and that our LLM-derived vectors are, moreover, able to capture these structures.

Individual Variability

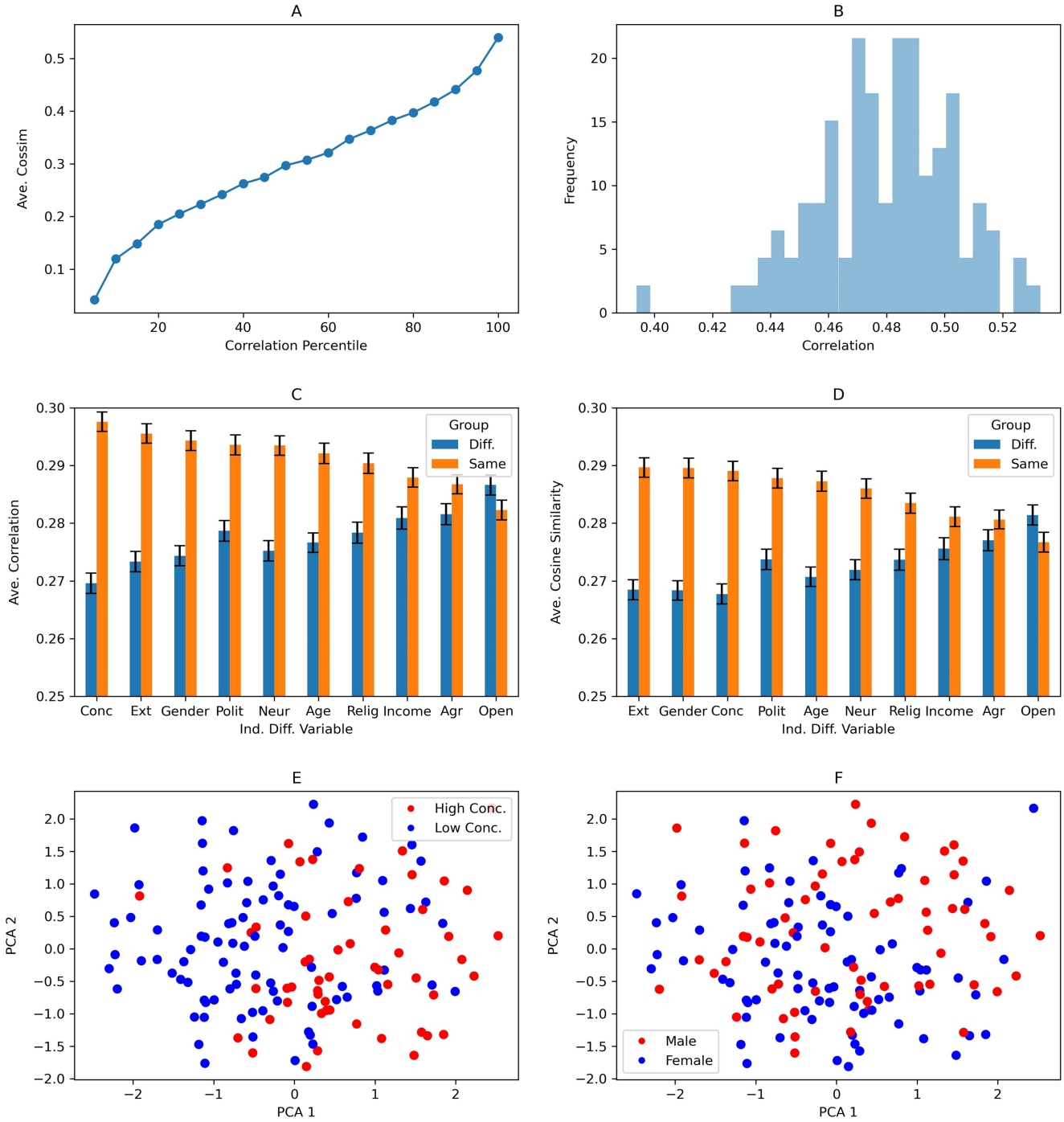
Next, we investigated the extent to which our approach could account for the structure of variability between individuals. Decision models describe people's preferences using attribute weights. Consequently, if our approach effectively captures individual differences, we would expect that similarities in attribute weights among pairs of individuals predict the degree to which the individuals engage in similar behaviors. To examine this, we estimated attribute weights for each individual by fitting our best performing model, SBERT-Items (with ridge penalty $a = 1$) on their complete ratings data set. Then we compared the cosine similarity of attribute weights with the correlation between people's ratings for each of the 10,440 unique pairs of participants. In other words, we tested whether $\mathrm{COR}(\boldsymbol{r}_j^p, \boldsymbol{r}_{j'}^p) \sim \mathrm{SIM}(\boldsymbol{w}_j, \boldsymbol{w}_{j'})$, where $\boldsymbol{w}_j$ is individual js 1,024-dimensional attribute weight vector, $\boldsymbol{r}_j^p$ is their 150-dimensional vector of ratings across all items, and j and j' are in range (1, 145) with $j \neq j'$. We found that the correlation between all pairs is .98 ($p < .001$, 95% CI = [0.98, 0.98]), showing that our approach captures the structure of variability between individuals almost perfectly. This is illustrated in Figure 4A. For clarity, we divided the 10,440 participant pairs into 20 equally sized bins, categorized based on the percentile ranking of their human correlation values. The figure displays the average cosine similarity of weight vectors within each bin. The findings depicted in Figure 4A reveal a clean positive relationship between participant correlations and the cosine similarities of their weight vectors.

Although the high correlation observed in the above paragraph would not be possible if the LLM vectors did not reflect the core attributes of the items, their high-dimensionality makes our analysis vulnerable to criticisms of over fitting. Thus, we performed a variant of the above test, in which we used weights recovered from half the items to predict between-individual correlations in responses on the other half of the items. In particular, we split the items into two portions randomly, obtained weights $\boldsymbol{w}_j$ from the first portion and ratings $\boldsymbol{r}_j^p$ from the second portion, and then tested whether $\mathrm{COR}(\boldsymbol{r}_j^p, \boldsymbol{r}_{j'}^p) \sim \mathrm{SIM}(\boldsymbol{w}_j, \boldsymbol{w}_{j'})$ for pairs of participants. We repeated these analysis 100 times to get a good estimate of the average correlation. This was found to be .48, $t(99) = 202.93$, $p < .001$, 95% CI = [0.478, 0.488]. The full distribution of correlations across the 100 random splits is shown in Figure 4B. These positive results provide definitive evidence that our decision model's weights do not overfit participants idiosyncrasies but rather capture participant heterogeneity in a robust and generalizable manner.

Next, we tested whether similarities and differences between different groups of individuals were captured by attribute weights, as would be predicted by a decision theoretic explanation for individual variability. For this, we performed binary splits on our full data set based on each demographic or psychographic variable collected in our study. For example, for gender, we divided our sample into male and female, for age into old and young (based on the median age of 33), for political affiliation into Democrat and Republican or Independent, for income into high and low earners (based on a threshold of $50,000), and for religion into religious and nonreligious. For each of the big five personality dimensions, we performed a median split on that dimension, dividing our sample into equally sized groups that are high or low on that dimension. For each of these groupings we calculated the average within and between group pairwise correlations in the ratings of pairs of individuals. Thus, for example, we measured the correlation between every man and every other man in our data set as well as between every woman and every other woman in our data set and aggregated these to get the average within-group correlation for gender. We also measured the correlation between every man and every woman in our data set and aggregated these to get the average between-group correlation for gender. We also did this for their associated attribute weights, to get the average within-group and average between-group cosine similarity of weights for the grouping. The averaged correlations and cosine similarities are shown in Figure 4C and 4D. If a given demographic or psychographic variable predicts risk-taking behavior, we would expect that people's ratings are more correlated if they belong to the same grouping on that variable than if they belong to different groupings on that variable. Indeed, we see this pattern for personality variables like conscientiousness and extraversion, as well demographic variables like gender. Importantly, our model captures this pattern by assigning similar attribute weights to individuals within the same groupings on these variables.

To better understand the nature of heterogeneity in attribute weights we further explored two of the above variables: conscientiousness and gender. Both have been implicated as predictors of risk taking in prior work, which has found that people who are similar in terms of these variables are also similar in terms of the behaviors they are most likely to do (Harris & Jenkins, 2006; Weber et al., 2002; Weller & Tikir, 2011). Indeed, in our own study we found that behaviors like working in a high stress environment and defending an unpopular issue that you believe in at a social occasion were rated highly by high conscientiousness participants whereas behaviors like not flossing teeth, not buying home insurance, and consuming marijuana were rated highly by low

Figure 4
Individual Variability



Note. (A) Average pairwise cosine similarity of participant weight vectors as a function of the pairwise correlation in participant ratings. Items are binned into 20 equally sized groups based on the participant correlations. (B) Distribution of out-of-sample correlations between weight vector cosine similarity and ratings correlations from 100 binary splits on the data. (C) Within and between group correlations in ratings across participants, for various demographic and psychographic variables. (D) Within and between group cosine similarities in the weights of participants, for various demographic and psychographic variables. (E) Scatter plot of first and second principle components of weight vectors, colored by conscientiousness grouping (high or low). (F) Scatter plot of first and second principle components of weight vectors, colored by gender grouping (male or female). Error bars indicate $\pm$ 1 *SE* in Panels A and B. Conc., Ext., Neur., Agr., and Open. refer to conscientiousness, extraversion, neuroticism, agreeableness, and openness to experience personality groupings, respectively, whereas Relig. and Polit. refer religious and political affiliation demographic groupings, respectively. Ave. Cossim = average cosine similarity; Diff. = difference; Conc = conscientiousness; Ext = extraversion; Polit = political; Neur = neuroticism; Relig = religious; Agr = agreeableness; Open = openness to experience; Ind. Diff. = individual difference; PCA = principle component. See the online article for the color version of this figure.

conscientiousness participants. Likewise investing in a speculative stock, bodybuilding, and having sex with a stranger were given much higher ratings by men than by women, whereas behaviors like eating an excessive amount of desserts and sweets, wearing provocative clothing, and ending a friendship were given much higher ratings by women than by men. Of course, these differences exist on the aggregate level and there is a large amount of variability within these groups as well. Figure 4E and 4F visualizes this variability in a scatter plot of individual-specific attribute weights on their first two principle components. Points are colored based on conscientiousness and gender in the two figures, respectively. Here we can see a fair amount of clustering, with most participants having closer decision weights to members of their own group. Roughly speaking, high (low) conscientiousness individuals occupy the bottom right (top left) of the plots whereas male (female) participants occupy the top right (bottom left) of the plots. Of course, there are also individuals who are more similar to members of other demographic and psychographic groups. This highlights the value of the continuous nature of our model's representations—Our approach is able to capture individual heterogeneity not by assigning people to a small number of discrete demographic or psychographic categories, but rather by capturing their complex multidimensional preferences in a graded and flexible manner.

Discussion

In this study we examined how well decision models based on LLM representations of behaviors predict the behaviors of participants (as measured by the DOSPERT inventory), as well as how well they describe variability across items and across individuals. We found high predictive accuracy rates, particularly for sentence vector models that consider contextual and syntactical information when representing the meaning of a sentence. Crucially, accuracy remained high even when using just 75%, 50%, or even 25% of the items, demonstrating no undue dependence on sample size. Furthermore, performance dropped to chance levels when LLM representations were replaced by random vectors, showing that our success relies on a meaningful mapping of behaviors into LLM vector spaces (rather than simply flexible high-dimensional vector spaces). Our approach also substantially outperformed an expert-coded model relying on DOSPERT labels for item domains. Taken together, these tests illustrate the predictive power of our framework, while ruling out potential limitations around scale and generalizability.

Our approach is also able to capture heterogeneity across items through its representations—behaviors that are rated similarly by human participants (e.g., behaviors that belong to the same domain) have similar attribute representations. Importantly, the graded nature of LLM representations allows them to capture item relationships that cannot be described by discrete experimenter-specified domains. This is why our models account for item heterogeneity within domain, for example, the fact that different health risks vary in terms of how correlated they are with each other. Additionally, our models capture variability across participants, as people who provide similar ratings to behaviors are given similar attribute weights. For this reason, our approach is also able to reproduce the effects of demographic and psychographic variables on behavior and account for participant heterogeneity within each grouping on our data set. In this way, our results show that the graded nature of LLM representations allows them to capture aspects of people's preferences that cannot be described using a small set of discrete demographic and psychographic variables.

One thing to note about the results in this section is that the best performing model was SBERT-Items, which used only LLM vectors for the behavior item's text (e.g., investing in a speculative stock) and not GPT-generated reasons for the behavior. This model did slightly better than SBERT-Reasons, which was based on LLM vectors for GPT-generated reasons. We suspect that this is the case because GPT generates reasons by itself relying on LLM representations of the item's texts. That is, the reasons that it generates do not provide any additional information that is not contained in and quantified by the rich (1,024 dimensional) vector representation of the item's text. For this reason, averaging over GPT's reasons leads to slightly worse performance. Of course, GPT-generated reasons are useful for interpreting the themes implicit in the model's (and in people's) representations of behaviors. What are these themes, and can they help us understand why people do or do not engage in certain behaviors? Examining this is the goal of Study 2.

Study 2

In Study 2 we wished to explore the reasons and attributes implicit in the representations of our models. For this purpose, we ran an extension of the experiment in Study 1 with a verbal protocol component (Ericsson & Simon, 1980; Schulte-Mecklenbeck et al., 2011; Weber et al., 2007) in which we asked participants to generate reasons for or against a subset of the behaviors in the extended DOSPERT inventory. Our goal was to compare the reasons generated by our models to those generated by human participants, and to furthermore interpret how the reason representations generated by our model capture item and individual heterogeneity in choice. A secondary goal was to test whether the high accuracy rates observed in Study 1 replicate in a new sample.

Experimental Method

Participants

We recruited 150 participants (79 female-identifying, 66 male-identifying, five other) from Prolific Academic, an online platform for conducting research with human subjects. The participants were aged between 18 and 70 years ($M = 37.39$, $SD = 12.25$) and were U.S. citizens fluent in English. All participants gave informed consent before taking part in the study. The study was approved by the UPenn Institutional Review Board (823184). We selected a sample size of 150 to be in line with the sample size of Study 1.

Materials and Procedure

Participants completed the extended DOSPERT scale using an identical interface to that in Study 1. After this, they were shown five randomly selected behaviors from the extended DOSPERT inventory and asked to list the reasons they choose to engage or not engage in that behavior. They were shown one behavior on each screen, were required to list at least one reason for that behavior, and could list up to 10 reasons in 10 separate text boxes shown on the screen. Participants were asked to write out their reasons in clear and complete sentences. After the verbal protocol task, participants completed the demographic and personality questionnaires used in

Study 1. Two participants failed the attention check embedded in the DOSPERT inventory and were removed from the study.

Transparency and Openness

We have reported how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. The analysis was performed in Python. The study design was preregistered at https://osf.io/amves/. Code and data are available at https://osf.io/3y6ku/.

Results

Replicating Study 1

Before analyzing reasons, we attempted to replicate the main findings of Study 1. As in Study 1, we found that the best performing model was SBERT-Items (which used vector representations of items' texts) with a ridge regression penalty $a = 1$. This model achieved an average LOOCV correlation between predicted and observed ratings of .42 ($SE = 0.01$). This was slightly better than the performance of SBERT-Reasons, which used vector representations of GPT-generated reasons for items. This model achieved a LOOCV correlation of .39 ($SE = 0.01$).

We also tested whether the pairwise correlation in the ratings of items was predicted by the pairwise cosine similarity of their LLM vectors, that is, whether $\mathrm{COR}(\boldsymbol{r}_i^b, \boldsymbol{r}_{i'}^b) \sim \mathrm{SIM}(\boldsymbol{x}_i, \boldsymbol{x}_{i'})$, where $\boldsymbol{x}_i$ is item is 1,024-dimensional vector and $\boldsymbol{r}_i^b$ is the 148-dimensional vector of ratings given to item I by individuals. We found that the correlation between cosine similarity and correlation pairs was .30 ($p < .001$, 95% CI = [0.28, 0.32]) for SBERT-Items and .25 for SBERT-Reasons ($p < .001$, 95% CI = [0.23, 0.27]). We also compared the cosine similarity of attribute weights with the correlation between people's ratings for each pair of participants, that is, whether $\mathrm{COR}(\boldsymbol{r}_j^p, \boldsymbol{r}_{j'}^p) \sim \mathrm{SIM}(\boldsymbol{w}_j, \boldsymbol{w}_{j'})$, where $\boldsymbol{w}_j$ is individual js 1,024-dimensional attribute weight vector and $\boldsymbol{r}_j^p$ is their 150-dimensional vector of ratings across all items. We found that the correlation between pairs was .98 for SBERT-Items ($p < .001$, 95% CI = [0.98, 0.98]) and .95 for SBERT-Reasons ($p < .001$, 95% CI = [0.95, 0.95]). These results show that the predictive performance of Study 1 persisted in Study 2, and moreover that our approach was able to capture the structure of variability between items and between individuals using its attribute vectors and weight vectors, respectively.

Human Versus Machine Reasons

Recall that we asked GPT to list five reasons for and five reasons against engaging in each behavior. This led to a total of 10 reasons per behavior and 1,500 reasons for the full inventory. Our verbal protocol analysis gave each behavior to an average of 4.93 participants, who generated an average of 2.17 reasons each. This resulted in an average of 10.69 unique human-generated reasons for each behavior and 1,582 reasons for the full inventory. Overall, the average number of characters used in each GPT-generated reason was 90.49 ($SD = 40.88$), whereas the average number of characters used in each human-generated reason was 44.50 ($SD = 30.45$), showing GPT generated substantially longer reasons than our participants.

We also examined the semantic content of the reasons generated by both humans and GPT. To do this, we used the 18 previously defined themes capturing key attribute dimensions within each DOSPERT domain (see Table 1). We coded the presence of each theme in each item by measuring the cosine similarity of theme vectors to the vectors for the item's GPT-generated reasons. Higher similarity indicates the theme is more represented in the reasons. As a first step, we calculated average theme frequencies across the full set of items to examine the prominence of the themes in our data. Additionally, we conducted the same analysis for participant-generated reasons to enable comparison with human data. These average theme frequencies are shown in Figure 5A, which indicates that themes like friendship and mental health are most common in both machine and human generated reasons, and that themes like adventure and education are less common in both machine and human generated reasons. Overall frequencies of the 18 themes are significantly correlated between humans and GPT ($r = .61, p < .01$). This suggests that GPT-generated reasons exhibit the same sensitivity to psychological factors implicitly shaping human explanations and reasons.

Interpreting Item Variability

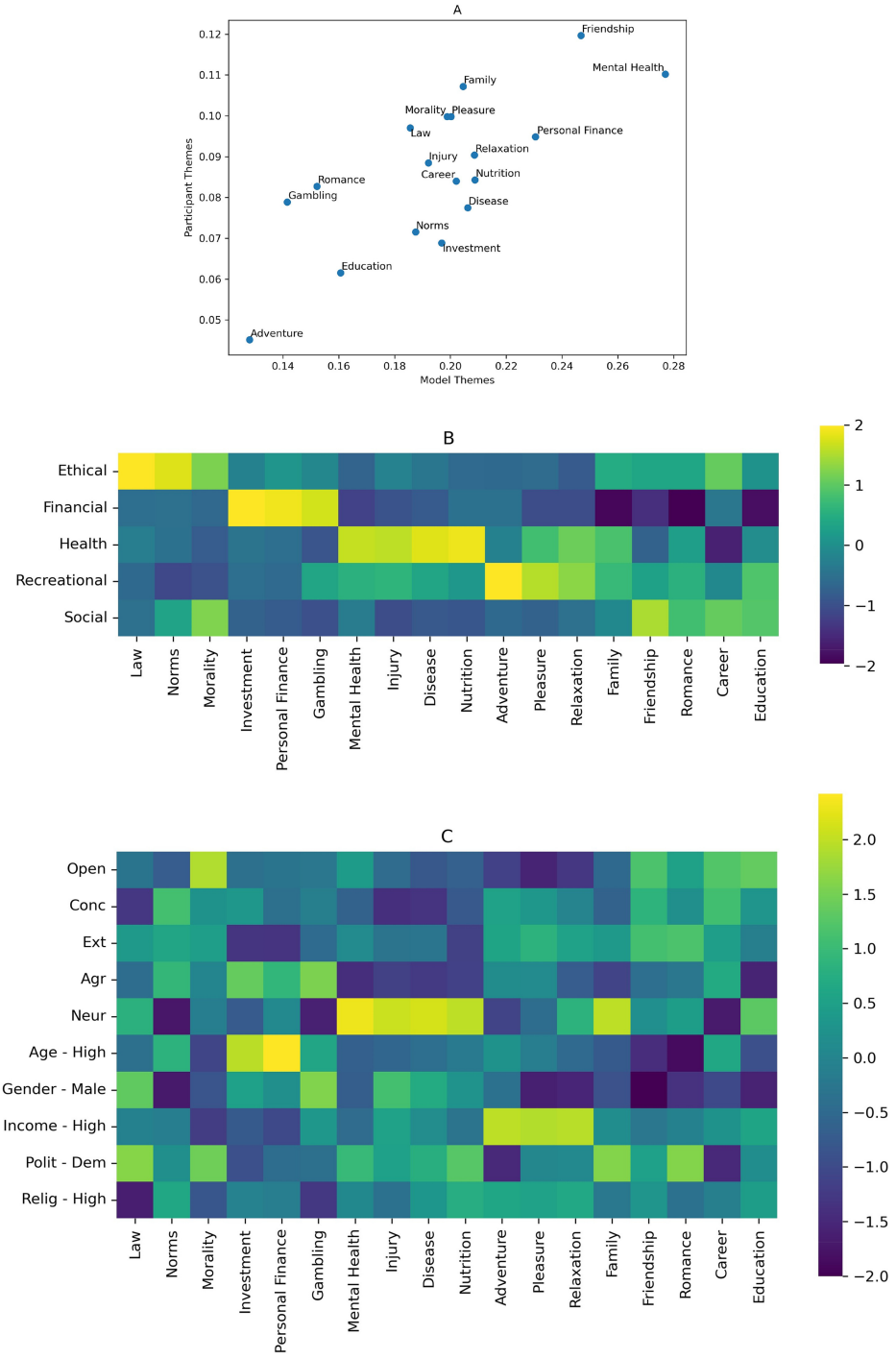
Our themes allow us to interpret sources of variability within and across DOSPERT domains. At a surface level, the five DOSPERT domains provide an intuitive taxonomy for distinguishing types of risk taking. However, they fail to fully explain rating patterns, as shown in Study 1. Why might items from the same domain elicit different ratings and why might items from different domains elicit similar ratings? We explore this in Figure 5B, which visualizes average (z-scored) theme frequencies by DOSPERT domain. This figure shows that although themes that are directly associated (and derived from) a given domain are most represented in that domain, they also systematically emerge in other types of domains. For instance, the morality theme has a high frequency of occurrence in reasons for social items, despite being derived from the ethical domain. Likewise, career and relaxation themes (original derived from social and recreational domains, respectively) have a high frequency in ethical and health reasons, respectively. Financial decisions are the most self-contained in terms of themes; however, cross-domain leakage exists there too—gambling themes, for example, frequently emerge in the context of recreational reasons. This analysis again shows that the experimenter-derived domain structure of DOSPERT is too coarse to fully capture the underlying complexity of DOSPERT, as items with similar interpretive themes may be rated similarly regardless of their surface domain label. This analysis also shows how LLM analysis can be useful for interpreting (and not just predicting) the sources of item variability in decision making.

Interpreting Individual Variability

We can apply a similar technique to interpret sources of individual variability in ratings. As in the analysis in Figure 3, we used binary splits on the 10 psychographic and demographic variables to divide participants into high and low groups for these variables. For each item, we then calculated mean ratings separately for each high and low group, and subtracted these to quantify the strength of association (and direction) of that item with that individual difference variable. This analysis showed, for example, that male participants rated items like investing in a speculative stock much higher than female participants. Finally, we correlated these item-level individual

Figure 5

Interpreting Themes at Play in Participant Ratings



Note. (A) Average theme frequencies in LLM-generated reasons (*x* axis) versus human-generated reasons (*y* axis) in Study 2. (B) Average theme frequencies in GPT-generated reasons as a function of the item's underlying DOSPERT domain. (C) Average theme frequencies in GPT-generated reasons for items given high versus low ratings in various psychographic and demographic splits of the data. Conc., Ext., Neur., Agr., and Open. refer to conscientiousness, extraversion, neuroticism, agreeableness, and openness to experience personality groupings, respectively, whereas Relig. and Polit. refer religious and political affiliation demographic groupings, respectively. Polit = political; Relig = religious; Open = openness to experience; Conc = conscientiousness; Ext = extraversion; Agr = agreeableness; Neur = neuroticism; Dem = democrat; LLM = large language model; GPT = generative pretrained transformer; DOSPERT = Domain-Specific Risk Taking scale. See the online article for the color version of this figure.

differences to our coded theme similarities, revealing which themes are most prominent in behaviors favored by people high or low on each psychographic or demographic split.

Figure 5C visualizes these correlations (after z-scoring). It shows systematic patterns in how different themes explain the behaviors of different groups of participants. For example, we see that high conscientiousness individuals tend to engage in behaviors with career risks, whereas low conscientiousness individuals engage in behaviors involving health and legal risks (explaining the patterns in Figure 4E). Likewise, men are more likely to engage in gambling, legal risks, and behaviors with the risk of injury, whereas women are more likely to engage in behaviors with social risks and leisure themes (explaining the patterns shown in Figure 4F).

As with Figure 5B, we also see why general domain labels may be ill-suited for understanding individual differences. For instance, law and norms (both of which are ethical themes) have differential relationships with individual difference variables. Norms are more reflected in reasons for the behaviors of high conscientiousness, high emotional stability (low neurotic), female, and high religiosity individuals, whereas legal themes are more reflected in reasons for the behaviors of low conscientiousness, highly neurotic, male, and low religiosity individuals. We also see similar disassociations for financial themes (with, e.g., personal finance and investment, but not gambling, being more common for older individuals), health themes (with, e.g., injury being more common for men but mental health being more common for women), recreational themes (with, e.g., adventure being more common for Republicans and relaxation being more common for high emotional stability individuals), and social themes (with, e.g., family being more common for high emotional stability individuals, friendship and romance being more common for high extraversion individuals, and friendship in particular being more common for women). Once again, this analysis demonstrates that the reasons that drive behavior vary in complex and nuanced ways both within and across risk domains. This explains why the approach outlined in this article effectively models individual differences: DOSPERT labels are simply too coarse, missing nuanced interactions, whereas LLM-derived vector representations capture rich attribute structures and describe how different groups disproportionally weigh those attributes.

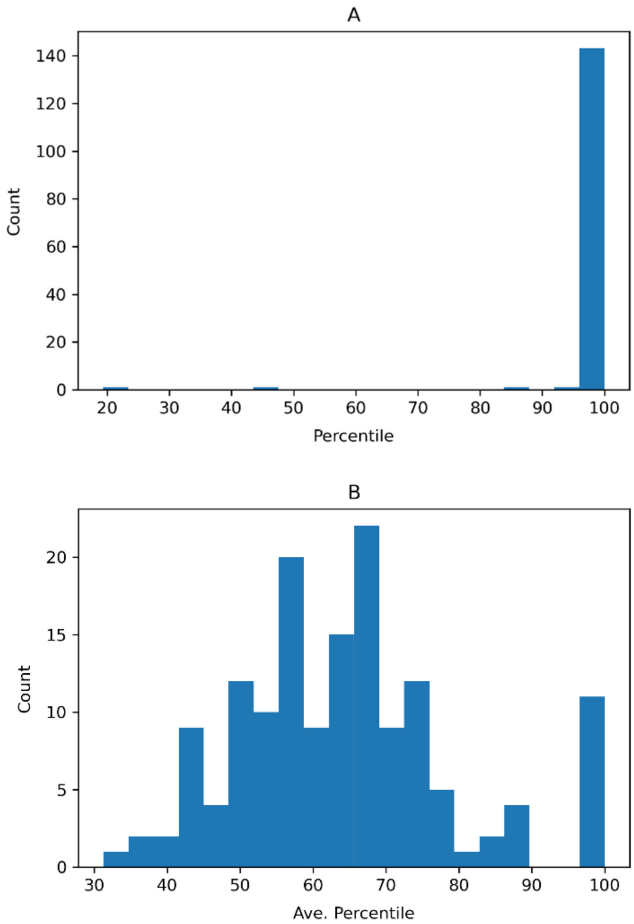
Predicting Reasons for Items

Our next test examined whether machine-generated reasons predicted human-generated reasons for the items. We did this by testing whether the reasons generated by GPT for a given item were more similar to the reasons that people generated for that item than they were to the reasons that people generated for other items. To quantify human and GPT-generated reasons, we passed each reason's text through SBERT to obtain its sentence vector. We then averaged the vectors for GPT-generated reasons for each item (as with the SBERT-Reasons model) as well as the vectors for human-generated reasons for each item. For item i we write these machine and human reason vectors as $\boldsymbol{x_i}$ and $\boldsymbol{y_i}$, respectively. Finally, we calculated the pairwise cosine similarity between each human and machine vector, and tested, for each i, whether SIM($\boldsymbol{x_i}$, $\boldsymbol{y_i}$) was larger than SIM($\boldsymbol{x_i}$, $\boldsymbol{y_{i'}}$) for $i' \neq i$. Statistically, this was done by calculating the percentile rank of SIM($\boldsymbol{x_i}$, $\boldsymbol{y_i}$) in the list, SIM($\boldsymbol{x_i}$, $\boldsymbol{y_1}$), SIM($\boldsymbol{x_i}$, $\boldsymbol{y_2}$) … SIM($\boldsymbol{x_i}$, $\boldsymbol{y_{150}}$). If machine generated reasons for behavior i are more similar to human generated reasons for that behavior than they are to human generated reasons for other behaviors, we would expect this percentile rank to be above 50%. In Figure 6A we show the distribution of these percentile ranks for each of the 150 items. Here we can see that the vast majority of percentile ranks are above 50%, with nearly all percentiles being very close to 100%. Overall, the average percentile is 98.61 which is significantly different to 50, $t(149) = 72.77$, $p < .001$, 95% CI = [97.28, 99.92]. This indicates that machine-generated reasons for an item are indeed more similar to human-generated reasons for that item than they are to human-generated reasons for other items.

Predicting Reasons for Individuals

We also examined whether our model could predict the reasons that each individual generated in the verbal protocol task. Recall that the

Figure 6
The Relationship Between Machine and Human-Generated Reasons



Note. (A) Percentile ranks showing how similar machine-generated reasons are to human-generated reasons for an item, relative to human-generated reasons for other items. (B) Percentile ranks showing how similar machine-generated reasons are to human-generated reasons for an item–individual pair, relative to reasons generated by other individuals for that item. We would expect percentile ranks of 50% if human and machine-generated were not systematically related. Ave. = average. See the online article for the color version of this figure.

item vectors, $\boldsymbol{x_i}$, represent the reasons that could be at play in a given behavior, whereas the individual-specific weights, $\boldsymbol{w_j}$, describe the participant's preferences over the dimensions of the reason space. Thus, a combination of $\boldsymbol{x_i}$ and $\boldsymbol{w_j}$ should uniquely reflect the reasons that a participant lists for engaging or not engaging in a given behavior, above and beyond the reasons that other participants list for that behavior. Again, if we quantify human-generated reasons by passing them through the SBERT model, and averaging them on the item level, then we would expect that the weighted attribute value on a given dimension should match up with the reason vector's value on that dimension. If we write the 1,024-dimensional vector for the reasons generated for item i by participant j as $\boldsymbol{z_{ij}}$ then we would expect that the kth dimension of $\boldsymbol{w_j}$ multiplied by the kth dimension of $\boldsymbol{x_i}$ should be roughly proportional to the kth dimension of $\boldsymbol{z_{ij}}$. Intuitively, if an individual j cares about dimension k (i.e., $\boldsymbol{w_j}$ has a high value on its kth dimension) and that dimension is relevant to behavior i (i.e., $\boldsymbol{x_i}$ has a high value on its kth dimension) then that dimension should be prominent in the vector representation of the participant's listed reasons ($\boldsymbol{z_{ij}}$ should have a high value on the kth dimension).

To test this rigorously, we attempted a variant of the analysis described in the previous section. For this we first calculated the element-wise product (also known as the Hadamard product) of $\boldsymbol{w_j}$ and $\boldsymbol{x_i}$, $\boldsymbol{w_j} \circ \boldsymbol{x_i}$ for each item and individual pairing. This measured the degree to which each dimension mattered for each item–individual combination. We then calculated the pairwise cosine similarity of $\boldsymbol{w_j} \circ \boldsymbol{x_i}$ with each $\boldsymbol{z_{ij}}$, and tested whether $\text{SIM}(\boldsymbol{w_j} \circ \boldsymbol{x_i}, \boldsymbol{z_{ij}})$ was larger than $\text{SIM}(\boldsymbol{w_j} \circ \boldsymbol{x_i}, \boldsymbol{z_{ij'}})$ for $j' \neq j$. Statistically, this was done by calculating the percentile rank of $\text{SIM}(\boldsymbol{w_j} \circ \boldsymbol{x_i}, \boldsymbol{z_{ij}})$ in the list [$\text{SIM}(\boldsymbol{w_j} \circ \boldsymbol{x_i}, \boldsymbol{z_{i1}})$, $\text{SIM}(\boldsymbol{w_j} \circ \boldsymbol{x_i}, \boldsymbol{z_{i2}})$ … $\text{SIM}(\boldsymbol{w_j} \circ \boldsymbol{x_i}, \boldsymbol{z_{i148}})$]. If our model is able to capture the reasons that individual j chooses to engage or not engage in behavior i over and above the reasons that other individuals list for that behavior i, then would expect this rank to be above 50%. In Figure 6B we show the distribution average percentile ranks for each of the 150 items. Here we can see that the majority of percentile ranks are above 50%, with the average percentile rank for the 150 items being 64.30. This is significantly different to 50, $t(149) = 11.52$, $p < .001$, 95% CI = [61.84, 66.75].

Discussion

Study 2 compared machine-generated reasons with those generated by human participants in a verbal protocol task. Participants were asked to list reasons why they do or do not engage in a given behavior. We used SBERT to transform these natural language reasons into high-dimensional vectors and attempted to predict them using vector representations of GPT-generated reasons. First, we showed that reasons generated by GPT captured key themes associated with the five original DOSPERT domains, explaining why items and individuals may vary in terms of their ratings. We also found that GPT-generated reasons for an item were more similar to human-generated reasons for that item than they were to human-generated reasons for other items. We additionally showed that best-fit attribute weights for participants uniquely predicted the reasons listed by those participants for items. Finally, Study 2 replicated the main results of Study 1 in a new sample of participants. Overall, the findings of Study 2 show that LLM-based decision models not only predict people's tendencies to engage in different behaviors, but can also predict (and, importantly, interpret) the reasons why different people chose to engage in different behaviors.

General Discussion

Summary

Decision models are powerful tools for describing the mental processes that give rise to behavior, and for explaining how behaviors vary across individuals and across items. However, until now, the use of decision models has been limited to highly stylized laboratory tasks which present participants with artificial stimuli defined on a small set of explicit attributes. This is due to the difficulty inherent in modeling the complex reasons that underlie real-world decision making. We attempted to solve this problem using LLMs, which have been shown to provide high quality quantitative representations of the meanings of sentences, including sentences used to describe common behaviors and their associated reasons (Brown et al., 2020; Devlin et al., 2018; Reimers & Gurevych, 2019). We tested our approach in two studies involving common risky behaviors. In Study 1, we administered an expanded version of the DOSPERT scale (Blais & Weber, 2006; Weber et al., 2002), consisting of 150 behaviors from five domains (finance, health, recreation, social, and ethical). Participants rated their propensity to engage in each behavior before answering several demographic and personality questionnaires. We used various LLM methods to obtain multiattribute representations for each of the behaviors, which we passed through a regularized linear regression to fit individual-specific decision weights. We evaluated our models by predicting out-of-sample ratings made by each individual for each item and found that our models achieved high accuracy rates, surpassing various benchmarks. Additionally, we found that the similarities of attribute vectors for items predicted the correlation in ratings between items and that the similarities of best-fit attribute weights of individuals predicted the correlation in ratings between individuals. We also found that our LLM-based behavior vectors captured the domain structure of risky choice (e.g., financial risks had vectors that were more similar to other financial risks than to recreational risks) and that best-fit attribute weights captured demographic and psychographic differences (e.g., men had weights that were more similar to other men than to women). These results demonstrate the predictive power of our approach and illustrate its ability to provide a decision theoretic foundation for psychometric analysis.

In Study 2, we ran a modification of the experiment in Study 1 with a verbal protocol component in which we asked participants to generate reasons for or against a subset of the behaviors in the extended DOSPERT inventory. First, we replicated the main results of Study 1. Then we used the prominence of 18 different themes in human and GPT-generated reasons to interpret the sources of variance in people's ratings. Finally, we tested whether the reasons generated by LLMs for a given item were more similar to the reasons that people generated for that item than they were to the reasons that people generated for other items. We also examined whether our model could predict the reasons that different individuals generated for a given item, based on their best-fit attribute weights. These tests showed that our approach achieved better predictions than by chance. Overall, the results of Study 2 show that our approach is not only useful for prediction, but can also be used to understand the reasons why people choose to engage or not engage in different behaviors.

Integrating Paradigms

We believe that the approach advanced in this article is particularly valuable, as it combines decision modeling and psychometric research traditions, which have different strengths and limitations. Decision modeling can provide an explanation for how people make decisions based on their preferences and decision processes, but it often requires simplifying assumptions and idealized scenarios. Psychometric research can provide a descriptive account of how real-world decision making varies across individuals, groups, items, and domains, but does not explain variance in decision making using theories of decision processes. By integrating these two traditions, our approach is able to provide a more comprehensive and realistic psychological account for the sources of variance in real-world decision making.

Of course, we are not the first to try to reconcile these two important paradigms. Prior applications of the DOSPERT inventory have used simple linear models to predict behavior from participant ratings of costs and benefits (Weber et al., 2002). Likewise, psychometric analysis of risk perception has derived multidimensional factor structures directly from people's ratings of experimenter-generated reasons and attributes (Fischhoff et al., 1978). These factors have been used in linear models to predict risk perceptions. Finally, recent work by Steiner et al. (2021) and Arslan et al. (2020) has elicited participant self-reports of reasons considered when evaluating general risk-taking tendencies. By manually coding these responses, and using the resulting codes in predictive analysis, these articles have investigated the psychological underpinnings of risk taking.

Our work builds on top of these important contributions by showing the power of LLMs for modeling complex risky decisions. Unlike prior approaches, the methods advanced in the current article do not need human ratings or self-reports to quantify the attributes at play in a given decision. This allows for our models to be applied in pure out-of-sample prediction. Moreover, LLM representations of behaviors, and preference weights on these representations derived from multiattribute decision models, are able to capture the complex web of social, legal, financial, and emotional considerations, at play in human behavior. This allows for more successful prediction and interpretation than using experimenter-derived or human-coded taxonomies of attribute structures.

Heterogeneous Representations Structures

A core property of the methodology in this article is its reliance on a singular representational structure when modeling subjective decisions. Specifically, all items are passed through the same LLM pipeline with the assumption that the resulting vectors describe the impressions, associations, and representations of all participants. In reality, people likely have idiosyncratic systems of associations reflecting unique backgrounds and experiences. This interindividual variability in semantic representation cannot be directly captured by applying a single invariant LLM transformation, which appears to be a fundamental constraint of our approach.

That said, the proposed approach of fitting individual-level weights on the LLM vectors does indirectly allow for heterogeneity in representational structures. Conceptually, allowing for customizable weighting across the semantic dimensions encoded in the vectors is directly analogous to participant-specific tweakings of the representational space itself. More concretely, imagine two individuals who associate the activity "drinking alcohol" with largely disjoint attributes due to their unique backgrounds. For Participant 1, drinking alcohol may cue biological effects like liver damage. For Participant 2 with personal experience managing alcoholics, attributes like impaired motor coordination resulting in accidents and falls may be more central. Our proposed model handles these divergent associative landscapes via the individual weight vectors learned by fitting individual-level ratings. Weights for Participant 1 effectively stretch alcohol-related dimensions of the space to align more closely with their personal associates, while those for Participant 2 independently elongate the coordination impairment dimensions. Thus, rather than concealing variation, the weighting mechanism implicitly allows the LLM vectors themselves to become participant-specific. Essentially, the flexible weights selectively prioritize distinct semantic aspects or reasons for each person, accommodating heterogeneity in unobserved representational structures. Indeed, this is the main reason why our approach is able to predict why different participants list different reasons for the same item, in Study 2.

It is worth noting that the property discussed above is closely related to transfer learning, which is responsible for the current successes of LLMs (see, e.g., Devlin et al., 2018 for a discussion). Transfer learning involves pretraining a base model on a generic objective (e.g., next word prediction) and then fine-tuning on small data sets for the target task (in our case, the participant's 150 ratings). The reason why transfer learning is so useful for LLMs is that fine-tuning allows the model to tweak its representational structure to match the target task without having to be fully trained on the target task, which generally does not have enough data to train a full language model from scratch. In our case, it is impossible to train an LLM on individual-level language data, and thus impossible to directly recover individual-specific representational structures for common behaviors. Given this limitation, our approach, which is an application of transfer learning to psychological problems, is the most practical and popular solution to the problem of heterogeneity in representational structure.

Dimensionality and Data Set Size

Another potential limitation involves the high dimensionality of the LLM representations and possible dependence on data set scale when fitting decision weights. While attribute vectors exceeding 1,000 dimensions are atypical for decision research, we believe this elevated complexity is essential for capturing real-world choice nuances—the very phenomenon this article seeks to model. Restricting attribute dimensions to two or three factors may be sufficient for simplified laboratory tasks with artificial objects, but it cannot encapsulate the full richness of organic judgment. Additionally, although it is true that model fitting requires more data than simple qualitative analysis, the 150 ratings used in this article is roughly in line with the number of observations used to fit individual-level decision models like cumulative prospect theory in laboratory-based psychological research (e.g., Glöckner & Pachur, 2012; He, Analytis, & Bhatia, 2022; He, Zhao, & Bhatia, 2022; Rieskamp, 2008). Finally, as demonstrated empirically, predictive performance remains high even when restricting training data to 25% of the original data, providing strong evidence of generalizability. Thus, rather than a weakness, the flexibility to represent fine-grained semantic distinctions through high dimensional representations is a strength of the proposed framework.

Interpretation

Another potential critique involves interpretability—by focusing predominately on predictive accuracy, the meaning of model parameters and attribute representations can be obscured. Additionally, on the surface, LLM-derived vectors appear to be inscrutable high-dimensional latent variables rather than discrete, interpretable characteristics or features. For this reason, they more likely reflect intuitive associations and impressions flowing from automatic cognition, rather than the clear high-level features used for deliberate reasoning. Although this is a fair point (and indeed, in prior work, we have argued that LLM representations are useful for modeling system 1 cognition—see Bhatia, 2017), LLM technologies nonetheless enable interpretation. We have shown this by coding items in terms of 18 themes reflected in GPT generated reasons. These themes illustrate why within-domain items can elicit dissimilar ratings and between-domain items can show similarities—uncovering cross-cutting psychological dimensions missed by existing DOSPERT taxonomies. Additionally, by correlating theme prominence with individual differences we are able to interpret the sources of variability across psychographic and demographic groups.

It is important to emphasize that the interpretative analysis in this article simply demonstrates how LLM-derived representations can be used to understand naturalistic decisions, and, by doing so, serve an explanatory role analogous to traditional decision science attributes. Other researchers could specify alternative themes tailored to their research questions and theoretical orientations, and reuse our pipeline to analyze those themes, without needing to collect any more data. In this way, our article shows how modern AI can transform qualitative psychological concepts into interpretable quantitative spaces amenable to rigorous analysis.

Psychological Mechanism

Some may question whether our approach truly provides insight into the psychological mechanisms underlying decision making, as the term "mechanism" can have different interpretations across subfields of psychology. For example, researchers focused on subconscious biases, heuristic shortcuts, or sequential information integration mechanisms might argue that our LLM-based weighted-additive decision models do not capture the type of nuanced processes they aim to study. However, we argue that our approach aligns well with the concept of psychological mechanism as employed in cognitive psychology, which traces its lineage back to the cognitive revolution (Miller, 2003; Simon, 1979). In this tradition, theories of mental processes emphasize the internal steps the mind employs to transform information into behavior. Our decision models, which instantiate these steps using a weighted additive rule applied to LLM-derived reasons, make a commitment to describing how people represent and manipulate information to generate responses and why they may differ in this process (this is in contrast to psychometric methods, which simply examine correlations between responses, and like previous behaviorist theories, are not explicitly concerned with how these responses are generated). Moreover, our analysis of decision models relies on standard methods used in the field, such as computationally specified algorithms whose parameters are fit to data (Study 1) and process data for evaluating the predictions of these algorithms (Study 2). Crucially, with these methods, we are able to provide a psychologically grounded answer to the question: "What are the sources of individual differences in risk taking?" Our answer is that different people care about different reasons. These differences take the form of different weights in a decision model and are reflected in the reasons people list when explicitly asked.

Now, it is worth noting that weighted additive decision models are not the only cognitive mechanisms possible. Indeed, contemporary research in cognitive psychology focuses on more sophisticated and complex information acquisition rules (Gigerenzer & Gaissmaier, 2011; Lee & Cummins, 2004; J. W. Payne et al., 1988; Roe et al., 2001; Zhao et al., 2022) that involve qualitatively different steps to simpler linear models (see, e.g., Jarecki et al., 2020 for a discussion of the differences between these classes of models). We believe that our approach is fully compatible with these contemporary theories. In fact, in Zhao et al. (2022), we have combined LLM representations of reasons with dynamic models of how these reasons are retrieved from memory and how retrieved reasons are aggregated to make decisions. We did not employ these complex models in the present study since the sources of individual differences are likely found in the reasons people prioritize rather than the nuanced operations they use to retrieve and aggregate these reasons. Methods like those in Zhao et al. are ideal for exploring the sequence of steps that people implement within a trial, but may be too complex for exploring differences in attribute weighting that manifest across trials and across individuals, which are better approximated by linear models. That said, we acknowledge that our approach is not the final word on naturalistic risky decision processes, and that the best mechanistic model would likely involve both complex information retrieval aggregation mechanisms (as in Zhao et al.) and individual-specific attribute weights (as in the current article). We believe that developing such a model should be the focus of future work.

Constraints on Generality

Our approach, though useful, has limitations. On an empirical level, we tested our approach in a sample of English-speaking U.S. participants. Since the LLMs in this article have been trained on extensive natural language data from this demographic, it is not particularly surprising that they can mimic the reasons and attributes considered by this demographic when making risky decisions. However, our successful predictions may not generalize to less represented demographics, particularly those with different cultural backgrounds and linguistic patterns, whose decisions can be shaped by a different set of norms, values, beliefs, and experiences. This underscores the need for more inclusive and diverse training data for LLMs, including multilingual data and data from nondigital (and possibly nontextual) sources. We expect that LLMs trained with these additional types of data should be able to provide an adequate account of the diverse rationales considered by a wide range of cultural and linguistic groups.

Complex Processes and Domains

There are also other limitations, in addition to the above constraints on generality. As discussed above, we use a linear model to represent decision making, which does not accommodate nuances like heuristic shortcuts and sequential reasoning. Future studies should integrate our approach with more complex decision models (e.g., Gigerenzer & Gaissmaier, 2011; Lee & Cummins, 2004; J. W. Payne et al., 1988; Roe et al., 2001; Zhao et al., 2022) to better understand how decision

variability relates to nuanced cognitive mechanisms. Such an analysis may be particularly useful for the study of ecological rationality, which examines the utility of different decision strategies for real-world cognition and behavior. Currently, most modeling work on decision strategies uses tightly controlled experiments, which precludes the assessment of performance in natural settings. The approach proposed in this article combines naturalistic elicitation and predictive modeling and can directly address this limitation, opening up new directions for the formal analysis of boundedly rational cognition and behavior.

Another imitation is that our work is solely focused on risk taking, which, while significant, is just one domain. Future research should apply our methods to different domains to broaden their applicability and verify the observed patterns. We suspect that our approach can be easily extended to both traditional decision-making scenarios such as intertemporal (He et al., 2023; Loewenstein & Prelec, 1992), social (Camerer, 2011; Fehr & Schmidt, 1999), or consumer (Green & Srinivasan, 1990; J. Payne et al., 1991) choice, as well as other topics currently studied using psychometric methods that rely on human ratings of lexical items. By converting the language used to describe common traits, attitudes, beliefs, and behaviors into vectors, we can also construct high-quality cognitive models capable of engaging in realistic deliberation and responding to psychological surveys in a human-like manner (see Abdurahman et al., 2024 for an additional example; for a detailed discussion, see Bhatia & Aka, 2022; also see Hofman et al., 2021; Yarkoni & Westfall, 2017 for discussions of the value of prediction in the social and behavioral sciences). In addition to providing better theories of human behavior, such an approach can also help in the development of AI agents that can better interact with people and respond to their behavior. Recent years have seen significant advancements in AI. If psychologists wish to contribute to the study of human behavior in this new era, they must leverage existing theories and combine them with innovative methods to address classic problems. This article demonstrates the feasibility of this research program, when applied to the study of individual differences in risk taking, and we are enthusiastic about the prospects that lie ahead.

References

Abdurahman, S., Vu, H., Zou, W., Ungar, L., & Bhatia, S. (2024). A deep learning approach to personality assessment: Generalizing across items and expanding the reach of survey-based research. *Journal of Personality and Social Psychology*, *126*(2), 312–331. https://doi.org/10.1037/pspp0000480

Aka, A., & Bhatia, S. (2022). Machine learning models for predicting, understanding, and influencing health perception. *Journal of the Association for Consumer Research*, *7*(2), 142–153. https://doi.org/10.1086/718456

Arslan, R. C., Brümmer, M., Dohmen, T., Drewelies, J., Hertwig, R., & Wagner, G. G. (2020). How people know their risk preference. *Scientific Reports*, *10*(1), Article 15365. https://doi.org/10.1038/s41598-020-72077-5

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review*, *124*(1), 1–20. https://doi.org/10.1037/rev0000047

Bhatia, S. (2019). Predicting risk perception: New insights from data science. *Management Science*, *65*(8), 3800–3823. https://doi.org/10.1287/mnsc.2018.3121

Bhatia, S., & Aka, A. (2022). Cognitive modeling with representations from large-scale digital data. *Current Directions in Psychological Science*, *31*(3), 207–214. https://doi.org/10.1177/09637214211068113

Bhatia, S., Loomes, G., & Read, D. (2021). Establishing the laws of preferential choice behavior. *Judgment and Decision Making*, *16*(6), 1324–1369. https://doi.org/10.1017/S1930297500008457

Bhatia, S., Olivola, C., Bhatia, N., & Ameen, A. (2022). Predicting leadership perception with large-scale natural language data. *Leadership Quarterly*, *33*(5), Article 101535. https://doi.org/10.1016/j.leaqua.2021.101535

Bhatia, S., & Richie, R. (2023). Transformer networks of human conceptual knowledge. *Psychological Review*, *131*(1), 271–306. https://doi.org/10.1037/rev0000319

Bhatia, S., Richie, R., & Zou, W. (2019). Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences*, *29*, 31–36. https://doi.org/10.1016/j.cobeha.2019.01.020

Bhatia, S., & Walasek, L. (2023). Predicting implicit attitudes with natural language data. *Proceedings of the National Academy of Sciences*, *120*(25), Article e2220726120. https://doi.org/10.1073/pnas.2220726120

Blais, A.-R., & Weber, E. U. (2006). A Domain-Specific Risk-Taking (DOSPERT) scale for adult populations. *Judgment and Decision Making*, *1*(1), 33–47. https://doi.org/10.1017/S1930297500000334

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., …, Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (pp. 1877–1901). Curran Associates.

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of Personality and Social Psychology*, *92*(5), 938–956. https://doi.org/10.1037/0022-3514.92.5.938

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive modeling*. Sage.

Busemeyer, J. R., Gluth, S., Rieskamp, J., & Turner, B. M. (2019). Cognitive and neural bases of multi-attribute, multi-alternative, value-based decisions. *Trends in Cognitive Sciences*, *23*(3), 251–263. https://doi.org/10.1016/j.tics.2018.12.003

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116–131. https://doi.org/10.1037/0022-3514.42.1.116

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Camerer, C. F. (2011). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pretraining of deep bidirectional transformers for language understanding* [Conference session]. Annual Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, Minnesota, United States (Vol. 1, pp. 4171–4186). https://aclanthology.org/N19-1423/

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, *9*(3), 522–550. https://doi.org/10.1111/j.1542-4774.2011.01015.x

Edwards, W. (1961). Behavioral decision theory. *Annual Review of Psychology*, *12*(1), 473–498. https://doi.org/10.1146/annurev.ps.12.020161.002353

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive-experiential and analytical-rational thinking styles. *Journal of Personality and Social Psychology*, *71*(2), 390–405. https://doi.org/10.1037/0022-3514.71.2.390

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*(3), 215–251. https://doi.org/10.1037/0033-295X.87.3.215

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868. https://doi.org/10.2139/ssrn.106228

Fischhoff, B., Slovic, P., Lichtenstein, S., Read, S., & Combs, B. (1978). How safe is safe enough? A psychometric study of attitudes towards

technological risks and benefits. *Policy Sciences*, *9*(2), 127–152. https://doi.org/10.1007/BF00143739

Frey, R., Duncan, S. M., & Weber, E. U. (2023). Towards a typology of risk preference: Four risk profiles describe two-thirds of individuals in a large sample of the U.S. population. *Journal of Risk and Uncertainty*, *66*(1), 1–17. https://doi.org/10.1007/s11166-022-09398-5

Frey, R., Pedroni, A., Mata, R., Rieskamp, J., & Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Science Advances*, *3*(10), Article e1701381. https://doi.org/10.1126/sciadv.1701381

Frey, R., Richter, D., Schupp, J., Hertwig, R., & Mata, R. (2021). Identifying robust correlates of risk preference: A systematic approach using specification curve analysis. *Journal of Personality and Social Psychology*, *120*(2), 538–557. https://doi.org/10.1037/pspp0000287

Gandhi, N., Zou, W., Meyer, C., Bhatia, S., & Walasek, L. (2022). Computational methods for predicting and understanding food judgment. *Psychological Science*, *33*(4), 579–594. https://doi.org/10.1177/0956797621104342 6

Garten, J., Hoover, J., Johnson, K. M., Boghrati, R., Iskiwitch, C., & Dehghani, M. (2018). Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior Research Methods*, *50*(1), 344–361. https://doi.org/10.3758/s13428-017-0875-9

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

Glöckner, A., & Pachur, T. (2012). Cognitive models of risky choice: Parameter stability and predictive accuracy of prospect theory. *Cognition*, *123*(1), 21–32. https://doi.org/10.1016/j.cognition.2011.12.002

Graham, J., Haidt, J., & Nosek, B. A. (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology*, *96*(5), 1029–1046. https://doi.org/10.1037/a0015141

Green, P. E., & Srinivasan, V. (1990). Conjoint analysis in marketing: New developments with implications for research and practice. *Journal of Marketing*, *54*(4), 3–19. https://doi.org/10.1177/002224299005400402

Harris, C. R., & Jenkins, M. (2006). Gender differences in risk assessment: Why do women take fewer risks than men? *Judgment and Decision Making*, *1*(1), 48–63. https://doi.org/10.1017/S1930297500000346

He, L., Analytis, P. P., & Bhatia, S. (2022). The wisdom of model crowds. *Management Science*, *68*(5), 3635–3659. https://doi.org/10.1287/mnsc.2021.4090

He, L., Wall, D., Reeck, C., & Bhatia, S. (2023). Information acquisition and decision strategies in intertemporal choice. *Cognitive Psychology*, *142*, Article 101562. https://doi.org/10.1016/j.cogpsych.2023.101562

He, L., Zhao, W., & Bhatia, S. (2022). An ontology of decision models. *Psychological Review*, *129*(1), 49–72. https://doi.org/10.1037/rev0000231

Highhouse, S., Nye, C. D., Zhang, D. C., & Rada, T. B. (2016). Structure of the DOSPERT: Is there evidence for a general risk factor? *Journal of Behavioral Decision Making*, *30*(2), 400–406. https://doi.org/10.1002/bdm.1953

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431–440. https://doi.org/10.1037/a0027373

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., Vazire, S., Vespignani, A., & Margetts, H. (2021). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), 181–188. https://doi.org/10.1038/s41586-021-03659-0

Howard, R. A. (1988). Decision analysis: Practice and promise. *Management Science*, *34*(6), 679–695. https://doi.org/10.1287/mnsc.34.6.679

Jarecki, J. B., Tan, J. H., & Jenny, M. A. (2020). A framework for building cognitive process models. *Psychonomic Bulletin & Review*, *27*(6), 1218–1229. https://doi.org/10.3758/s13423-020-01747-2

Josef, A. K., Richter, D., Samanez-Larkin, G. R., Wagner, G. G., Hertwig, R., & Mata, R. (2016). Stability and change in risk-taking propensity across the adult life span. *Journal of Personality and Social Psychology*, *111*(3), 430–450. https://doi.org/10.1037/pspp0000090

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, *47*(2), 263–292. https://doi.org/10.2307/1914185

Keeney, R. L., & Raiffa, H. (1993). *Decisions with multiple objectives: Preferences and value trade-offs*. Cambridge University Press.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Lauriola, M., Levin, I. P., & Hart, S. S. (2007). Common and distinct factors in decision making under ambiguity and risk: A psychometric study of individual differences. *Organizational Behavior and Human Decision Processes*, *104*(2), 130–149. https://doi.org/10.1016/j.obhdp.2007.04.001

Lee, M. D., & Cummins, T. D. (2004). Evidence accumulation in decision making: Unifying the "take the best" and the "rational" models. *Psychonomic Bulletin & Review*, *11*(2), 343–352. https://doi.org/10.3758/bf03196581

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *Roberta: A robustly optimized bert pretraining approach*. arXiv. https://arxiv.org/abs/1907.11692

Loewenstein, G., & Prelec, D. (1992). Anomalies in intertemporal choice: Evidence and an interpretation. *The Quarterly Journal of Economics*, *107*(2), 573–597. https://doi.org/10.2307/2118482

Lu, H., Wu, Y. N., & Holyoak, K. J. (2019). Emergence of analogy from relation learning. *Proceedings of the National Academy of Sciences*, *116*(10), 4176–4181. https://doi.org/10.1073/pnas.1814779116

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78. https://doi.org/10.1016/j.jml.2016.04.001

Markowitz, H. M. (1959). *Portfolio selection*. Wiley.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (pp. 3111–3119). Curran Associates.

Miller, G. A. (2003). The cognitive revolution: A historical perspective. *Trends in Cognitive Sciences*, *7*(3), 141–144. https://doi.org/10.1016/S1364-6613(03)00029-9

Payne, J., Bettman, J. R., & Johnson, E. J. (1991). Consumer decision making. In T. S. Robertson & H. H. Kassarjian (Eds.), *Handbook of consumer behaviour* (pp. 50–84). Prentice-Hall.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1988). Adaptive strategy selection in decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(3), 534–552. https://doi.org/10.1037/0278-7393.14.3.534

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science*, *17*(5), 407–413. https://doi.org/10.1111/j.1467-9280.2006.01720.x

Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality*, *41*(1), 203–212. https://doi.org/10.1016/j.jrp.2006.02.001

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks* [Conference session]. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Hong Kong. https://aclanthology.org/D19-1410/

Richie, R., Aka, A., & Bhatia, S. (2023). Free association in a neural network. *Psychological Review*, *130*(5), 1360–1382. https://doi.org/10.1037/rev0000396

Richie, R., Zou, W., & Bhatia, S. (2019). Predicting high-level human judgment across diverse behavioral domains. *Collabra: Psychology*, *5*(1), Article 50. https://doi.org/10.1525/collabra.282

Rick, S. I., Cryder, C. E., & Loewenstein, G. (2008). Tightwads and spendthrifts. *Journal of Consumer Research*, *34*(6), 767–782. https://doi.org/10.1086/523285

Rieskamp, J. (2008). The probabilistic nature of preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(6), 1446–1465. https://doi.org/10.1037/a0013646

Roe, R. M., Busemeyer, J. R., & Townsend, J. T. (2001). Multialternative decision field theory: A dynamic connectionst model of decision making. *Psychological Review*, *108*(2), 370–392. https://doi.org/10.1037/0033-295X.108.2.370

Rushton, J. P., Chrisjohn, R. D., & Fekken, G. C. (1981). The altruistic personality and the Self-Report Altruism Scale. *Personality and Individual Differences*, *2*(4), 293–302. https://doi.org/10.1016/0191-8869(81)90084-2

Schulte-Mecklenbeck, M., Kühberger, A., & Johnson, J. G. (2011). *A handbook of process tracing methods for decision research: A critical review and user's guide*. Psychology Press.

Schwartz, B., Ward, A., Monterosso, J., Lyubomirsky, S., White, K., & Lehman, D. R. (2002). Maximizing versus satisficing: Happiness is a matter of choice. *Journal of Personality and Social Psychology*, *83*(5), 1178–1197. https://doi.org/10.1037/0022-3514.83.5.1178

Simon, H. A. (1979). *Models of thought*. Yale University Press.

Singh, M., Richie, R., & Bhatia, S. (2022). Representing and predicting everyday behavior. *Computational Brain & Behavior*, *5*(1), 1–21. https://doi.org/10.1007/s42113-021-00121-2

Slovic, P. (1987). Perception of risk. *Science*, *236*(4799), 280–285. https://doi.org/10.1126/science.3563507

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1977). Behavioral decision theory. *Annual Review of Psychology*, *28*(1), 1–39. https://doi.org/10.1146/annurev.ps.28.020177.000245

Steiner, M. D., Seitz, F. I., & Frey, R. (2021). Through the window of my mind: Mapping information integration and the cognitive representations underlying self-reported risk preference. *Decision*, *8*(2), 97–122. https://doi.org/10.1037/dec0000127

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Tian, K. T., Bearden, W. O., & Hunter, G. L. (2001). Consumers' need for uniqueness: Scale development and validation. *Journal of Consumer Research*, *28*(1), 50–66. https://doi.org/10.1086/321947

Weber, E. U., Blais, A., & Betz, N. E. (2002). A domain-specific risk-attitude scale: Measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, *15*(4), 263–290. https://doi.org/10.1002/bdm.414

Weber, E. U., & Johnson, E. J. (2009). Mindful judgment and decision making. *Annual Review of Psychology*, *60*(1), 53–85. https://doi.org/10.1146/annurev.psych.60.110707.163633

Weber, E. U., Johnson, E. J., Milch, K. F., Chang, H., Brodscholl, J. C., & Goldstein, D. G. (2007). Asymmetric discounting in intertemporal choice: A query-theory account. *Psychological Science*, *18*(6), 516–523. https://doi.org/10.1111/j.1467-9280.2007.01932.x

Weller, J. A., & Tikir, A. (2011). Predicting domain-specific risk taking with the HEXACO personality structure. *Journal of Behavioral Decision Making*, *24*(2), 180–201. https://doi.org/10.1002/bdm.677

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zhao, W., Richie, R., & Bhatia, S. (2022). Process and content in decisions from memory. *Psychological Review*, *129*(1), 73–106. https://doi.org/10.1037/rev0000318

Received June 29, 2023
Revision received March 20, 2024
Accepted April 1, 2024 ■