Decision

Verbal Reports as Data Revisited: Using Natural Language Models to Validate Cognitive Models

Tehilla Ostrovsky and Ben R. Newell

Online First Publication, August 22, 2024. https://dx.doi.org/10.1037/dec0000243

CITATION

Ostrovsky, T., & Newell, B. R. (2024). Verbal reports as data revisited: Using natural language models to validate cognitive models.. *Decision*. Advance online publication. https://dx.doi.org/10.1037/dec0000243

AMERICAN PSYCHOLOGICAL ASSOCIATION

Decision

© 2024 American Psychological Association
ISSN: 2325-9965

https://doi.org/10.1037/dec0000243

Verbal Reports as Data Revisited: Using Natural Language Models to Validate Cognitive Models

Tehilla Ostrovsky[1, 2] and Ben R. Newell[1, 3]

[1] School of Psychology, University of New South Wales Sydney
[2] Department of Psychology, Ludwig Maximilian University
[3] Institute for Climate Risk and Response, University of New South Wales Sydney

We propose a novel technique that uses individuals' verbal reports to validate psychological processes assumed by computational cognitive models. We capitalize on recent advances in natural language processing models, especially their context-sensitivity, and recommend using them to classify participants' unstructured verbal descriptions of the strategies they use to perform tasks. This framework emphasizes that verbal descriptions are a valuable and under-utilized source of data for ensuring cognitive models align with their psychological assumptions. We argue that this framework can encompass a broad range of cognitive tasks, including problem solving, reasoning, memory, categorization, and decision making.

Keywords: natural language processing models, self-report, verbal description, decision making, cognitive model validity

Computational approaches to understanding cognition are commonplace. Advances in, and the availability of, resources for modeling have led to a huge increase in the number and variety of cognitive models. Many of these models are able to accurately describe people's behavior across a wide range of tasks and contexts. Despite their accuracy, a debate remains over the extent to which such models should be considered as capturing something fundamental about underlying psychological processes (e.g., Szollosi et al., 2023). Do such models directly mimic these processes or simply provide mathematical tools for examining them? This debate has yielded several fruitful avenues of investigation but to date has largely neglected one potentially important source of data that we believe could help bridge the gap between using models as tools versus explanations of behavior. This additional data source is participants' verbal descriptions of how they perform tasks.

The idea that participants can explain their own behavior reliably has received short shrift in some corners of the literature (most famously by Nisbett & Wilson, 1977). However, more recent re-evaluations of this perspective have made a case for how and why careful elicitation of participants' verbal descriptions can inform our understanding of psychological processes (e.g., Newell & Shanks, 2014, 2023; Szollosi & Newell, 2020). Seminal research by Ericsson, Simon, and others explored how self-report questionnaires and unstructured verbal reports provide potentially rich sources of data that can be used to build psychological explanations of behavior (Ericsson & Simon, 1980; Fox et al., 2011; Ranyard & Svenson, 2011; Svenson, 1979, 1989). The innovation in our work is to show how these data can also be used to help validate the psychological assumptions inherent in our cognitive and computational models of behavior.

Tehilla Ostrovsky https://orcid.org/0000-0002-8875-5832

This research was funded by an Australian Research Council Discovery Project Grant (DP190101076).

Tehilla Ostrovsky played a lead role in conceptualization, formal analysis, methodology, writing–original draft, and writing–review and editing. Ben R. Newell played a supporting role in conceptualization, supervision, and writing–review and editing.

Correspondence concerning this article should be addressed to Tehilla Ostrovsky, Department of Psychology, Ludwig Maximilian University, Akademiestrasse 7, 80799 Munich, Germany. Email: tehilla.mo@gmail.com or tehila.mechera-ostrovsky@psy.lmu.de

In essence, we suggest that verbal report data should be used in the way that other sources of data are often used to inform and constrain cognitive models such as data from neuroimaging (Carter et al., 2010), process-tracing, and eye-tracking (Gluth et al., 2020; Krajbich, 2019; Krajbich & Rangel, 2011; Vanunu et al., 2021).

Our proposed framework enables the validation of assumptions inherent in cognitive models using open-end, unstructured text-based responses. This approach uses the power of natural language processing (henceforth NLP) models to evaluate participants' unstructured self-reported strategies and to classify them according to their content. This classification represents a new source of information that allows a more refined insight into two important aspects of experimental work: (1) individuals' knowledge and understanding of the experimental task and (2) the strategies participants think they used to solve or perform a particular task. The rapid development of NLP models and their outstanding performance across a wide range of text-based tasks allows this source of data to provide stronger support (or challenge) to the formal descriptions of human thought and behavior made by cognitive models. Hence, it serves in helping to hold cognitive models accountable for their psychological assumptions.

Prior to outlining our framework, it is important to highlight how our approach sets itself apart from the existing literature linking NLPs with human psychology. One focus of this literature is on the similarities or differences between human reasoning and that of NLPs Bhatia (2022), Bhatia and Richie (2022), Bhatia and Stewart (2018), and Binz and Schulz (2022, 2023). A second focus is to use NLP models to help generate experimental stimuli, enabling researchers to acquire naturalistic stimuli while retaining the necessary level of experimental control during lab-based tasks. (Stimuli, which, in turn, can be used to construct and inform computational models of cognition; e.g., Bhatia, 2022; Bhatia & Richie, 2022).

Our objective is different to both of these applications. We do not seek to compare the "cognitive abilities" of NLPs with those of humans, nor evaluate the significance of their similarities; instead, we aim to revitalize the elicitation of verbal reports and their direct comparison with the constructs (i.e., parameters) defined in our cognitive models. By leveraging the automated capabilities of NLPs to efficiently handle text-based data, we aim to contrast these analyses with the parameters of cognitive models that are interpreted as holding psychological significance.

Theoretical Framework

The basic principle of cognitive models is the necessity to formally describe human reasoning and behavior using abstract mathematical specifications (e.g., Navarro, 2021; Proulx & Morey, 2021; Singmann et al., 2022). The objective is to achieve high levels of precision while eliminating any subjective interpretation about the underlying processes. Although using ordinary language to describe cognitive processes is clearly a challenge to this principle, mathematical specifications are yet just another, very precise verbal description of those processes.

To illustrate this idea consider the standard class of evidence accumulation models (Brown & Heathcote, 2008; Ratcliff, 1978; Ratcliff et al., 2016). In these models, noisy evidence starts accumulating at a point between two response boundaries (also known as a "decision threshold") and terminates when one of the boundaries is reached. Let us consider a very simple lab-based task, a choice between two food items, an apple and a pear. Let us further consider Person A, who has a clear preference for apples over pears. An evidence accumulation model will be able to capture a bias toward an apple based on Person A's reaction time distributions and will use either a biased starting point or a lower boundary for the preferred option (Ratcliff et al., 2016).

Now, imagine asking Person A to engage in describing their behavior while making their decision. Person A might describe her preferences for the apple using the features important to her (e.g., taste, texture and other nutritional facts). We can further expect that when explicitly asked, Person A will provide her reasoning for her strong preference for apples describing this choice as clear, obvious or quick due to her weak interest in pears. Evidence accumulation models help capturing an initial bias toward the apple with a precise numerical value that best describes the empirical data provided by Person A. However, it is also clear that if this parameter indeed captures the process of an underlying behavior precisely (and simply), it can also be explained by a lay person using natural language in a similar way to those described in the "biased" starting point (or by a lower decision boundary).

Following the example, derived from their text response, Person A indeed has an initial bias toward one of the options, which can validate an introduction of a bias to a computational model. This might contrast with a situation in which a participant states that they did not notice or recognize the pear and so chose the apple. In this case, a model with a biased starting point seems inappropriate for capturing the thought process engaged in by the participant.

This simple example helps reveal how verbal reports can be beneficial for validating and supporting the assumptions underlying our models, particularly in terms of their parameters. However, this process is not straightforward and requires careful planning. Earlier work by Ericsson and Simon (1980), sheds light on the challenges in this area and remains relevant even with the advancements of contemporary NLP models.

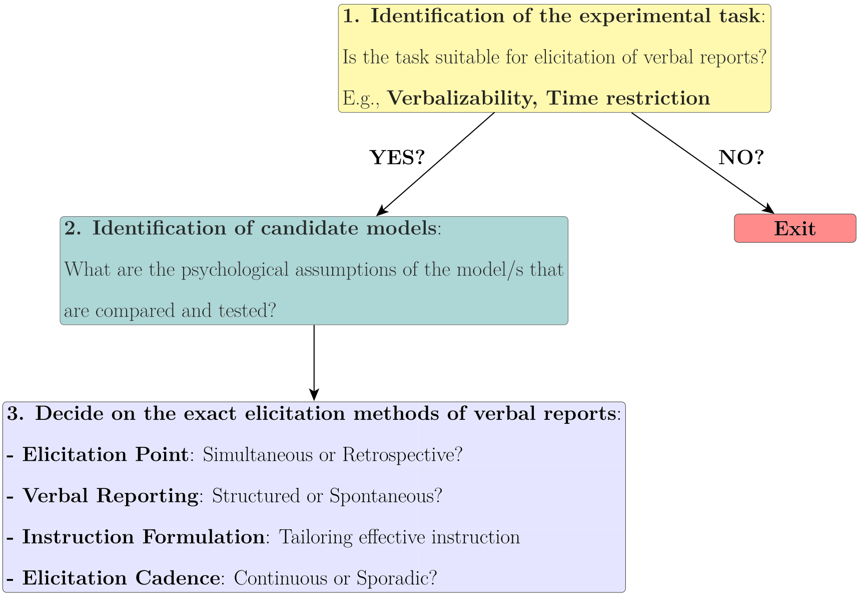
The Legacy of Ericsson and Simon

The seminal work by Ericsson and Simon (1980) explores in detail the utility of verbal reports as data in experimental settings. We review some of the key aspects here to illustrate their relevance for our framework. Figure 1 situates the aspects into a flowchart.

The first, and perhaps self-evident, step is to consider whether the primary task of interest is amenable to the elicitation of verbal reports. This criterion of "verbalizability" (Step 1, Figure 1) refers to the requirement that information regarding the utilized strategy must already exist

Figure 1

Flowchart Illustrating the Steps for Eliciting Verbal Reports


1. Identification of the experimental task:
Is the task suitable for elicitation of verbal reports?
E.g., **Verbalizability, Time restriction**
YES?
NO?
Exit
2. Identification of candidate models:
What are the psychological assumptions of the model/s that are compared and tested?
3. Decide on the exact elicitation methods of verbal reports:
- **Elicitation Point**: Simultaneous or Retrospective?
- **Verbal Reporting**: Structured or Spontaneous?
- **Instruction Formulation**: Tailoring effective instruction
- **Elicitation Cadence**: Continuous or Sporadic?

Note. *Identify the Experimental Task*: Start by assessing the experimental task's suitability for eliciting verbal reports. Analyze factors such as *verbalizability* and *time constraints identify candidate models*: Once the task is defined, select cognitive models for comparison and testing. Describe their psychological assumptions, akin to verbal reports and determine if these assumptions can be expressed verbally. *Decide on Elicitation Methods*: Make crucial decisions about collecting verbal reports during the experiment. Consider: *Elicitation Point*: Choose between concurrent or retrospective collection of verbal reports. *Verbal Reporting*: Choose between providing structured and directed guidance on the focus areas for the verbal reports and permitting participants to freely express their thoughts in their accounts, *Instruction Formulation*: Determine how to strategically craft instructions that prompt participants to articulate their actual strategies. *Elicitation Cadence*: Decide on whether to gather reports consistently after each task trial or intermittently at predetermined intervals. See the online article for the color version of this figure.

in a verbal form for participants to effectively report it, and that the reporting of that information does not negatively disrupt primary task performance. Not all experimental designs will accommodate this criterion. For instance, studies investigating perceptual-motor functions, have shown that verbalization can exert an influence on how tasks are executed (e.g., Beilock & Carr, 2005). Similarly, some speed–accuracy tasks may be less suitable for this method due to the challenges associated with direct articulation in time-restricted environments (e.g., Nisbett & Wilson, 1977). In each case, it is important to consider the potential consequences of asking participants to articulate information that is not naturally processed verbally or is challenging to express in words. Care must be taken to ensure that this process in itself does not alter the way in which participants approach a task. One way this can be done is by comparing performance on a primary task either with or without concurrent verbalization/strategy questioning (e.g., Lagnado et al., 2006).

If amenability to verbalization is established, the next step is to consider the candidate cognitive or computational models that can be applied to the primary task (Step 2, Figure 1). Here, the important aspect is that the psychological assumptions underpinning particular parameters in a given model have the potential to be captured in verbal reports via natural language. As our earlier example with fruit choice illustrated, the idea is to be able to map what people say on to psychologically significant parameters in the cognitive model (we expand further on this process—which is core to our approach—in the next section). If this kind of mapping is achievable, then researchers can proceed to the third step.

Step 3 requires detailed consideration of the exact method for eliciting verbal reports. Here, researchers need to decide whether to elicit reports via *Concurrent Probing*, which involves questioning subjects during a task about their underlying hypotheses or *Retrospective Verbalization*, in which participants are queried posttask, either immediately or after several trials. Additional considerations are whether elicitation is *directed* or *undirected*. Directed probes present specific queries to participants, aiming to discern whether they employed a particular strategy or approach. Conversely, undirected probes grant participants the latitude to freely elucidate their thought processes. Figure 1 provides additional information on other methodological choices.

Linking Verbal Reports With Cognitive Models via Language Models

Successful navigation of the steps outlined in Figure 1, brings a researcher to the point of having a candidate task that is amenable to verbalization, a set of candidate cognitive models for explaining performance on that task, and consensus on the best methods for eliciting verbal reports. Figure 2 presents the conceptual framework underlying the next stages in linking the content of the verbal reports to the psychological assumptions underlying the parameters in a cognitive model, via NLP models.

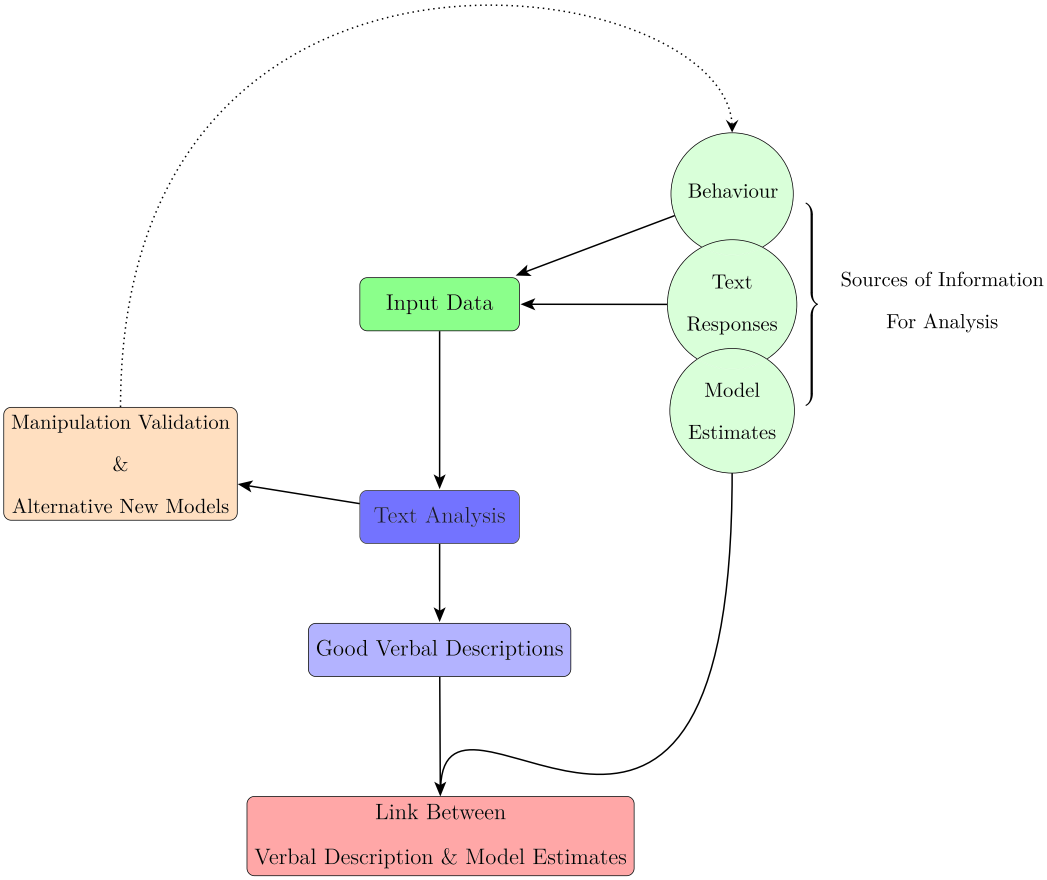
At this stage the researcher conducts the experiment to examine a specific prediction (to test a theory) and collects both behavioral and text-based data to test it. At the conclusion of the experiment, the researcher will have three sources of information for analysis: primary task behavior (e.g., trial by trial responses), verbal reports (whether obtained concurrently with the experiment, immediately afterwards, or at intermittent intervals throughout the study), and model estimates obtained from using empirical data to fit cognitive models (at both individual and group levels, potentially; See Figure 2, circular nodes).

The crucial next stage is the text analysis (blue rectangle, Figure 2) in which verbal reports are used as input to the natural language models. The benefit here is that natural language models are designed to scan and analyze human-generated (nonconstrained) text. They are based on supervised machine learning algorithms that have been trained on vast amounts of text data in order to learn particular functions. A recent influential development was the introduction of an "attention mechanism" which allows the embedding of contextual meaning into words based on their location in a sentence (Vaswani et al., 2017).[1]

[1] These models start by processing every word sequentially—similar to earlier models of word embeddings (i.e., vector representation for words, where similar words have similar embeddings/vector representations). In this class of models, each word has a unique embedding. It then assigns a weight to each word embedding that represents the position of that word in the text (e.g., in a sentence). In a last step, a weight is generated for each word that represents how much its meaning or grammatical form (e.g., nouns, verbs, adjectives, adverbs, prepositions, etc.) is dependent on other words in the sentence. For instance, the sentence "I took my dog for a walk," the subject of the sentence are given more power relative to "my" (which is possessive pronoun.) or "walked" which is the verb of the sentence.

Figure 2

Our Conceptual Framework for the Link Between Verbal Description Analysis and Cognitive Models


Behaviour
Text Responses
Model Estimates
Sources of Information For Analysis
Input Data
Manipulation Validation & Alternative New Models
Text Analysis
Good Verbal Descriptions
Link Between Verbal Description & Model Estimates

Note. The process starts with input data consisting of three (correlated) sources of information (green circles). To identify good verbal descriptions, the verbal reports are evaluated via text analysis models and are matched with behavior (blue rectangles). Once text analysis is performed it provides both validation on participants' understanding of the experimental manipulation and new ideas for potential models (orange rectangle). Finally, the link between the individual NLP model outputs and the individual model fits are correlated (red rectangle). NLP = natural language processing. See the online article for the color version of this figure.

This improvement was made possible due to the existence of a class of neural network models called "transformers" (Devlin et al., 2018). With these new mechanisms, NLP models improved their ability to make sense of human-generated text and were able to outperform the old models across a large number of text-based tasks (Devlin et al., 2018; Wolf et al., 2020). This feature is particularly useful in the analysis of verbal report data in which descriptions of strategies may remain relatively impoverished despite attempts to guide participants' responses.

A thorough analysis of verbal reports can take advantage of NLP techniques, including methods spanning from vectorization and labeling, to clustering and keyword extraction. By transforming qualitative verbal data into quantitative, structured information, researchers can attempt to effectively discern the underlying cognitive mechanisms at play during experimental tasks.

For example, one way to classify participants' verbal descriptions, is to allow language models to test the relation between a set of candidate labels and then select the one that best describes a given report. These candidate labels are provided to the model by the experimenter based on assumptions about the different kinds of behaviors participants might exhibit in performing the task. In practice, the language model takes a sentence from a verbal report and a list of candidate labels as input and returns accuracy score(s) to each tested label(s) (Sanh et al., 2022; Yin et al., 2019). This "accuracy" score ranges between 0 and 1, which can be understood as the "confidence" the model assigns to the ability of a label to describe the sentences provided by participants in their verbal reports.

Ideally, participant-generated verbal reports will consist of two important elements: (1) A description of *how* the problem was solved or choice was made; these descriptions represent the "algorithms" that generated the target behavior and (2) the reason (*why*) a specific algorithm was used. In reality, the reports might not capture these aspects clearly, requiring steps of refinement to what we term "good verbal descriptions" in our framework. The main goal in this step of the analysis is to correlate these label scores with the individual cognitive model fits (that offer the computational "description" of how the task was performed) to assess the validity, or psychological plausibility of the cognitive models' core assumptions. To assess the quality of the verbal descriptions requires comparing the selected (i.e., most suitable) label of an individual's strategy to their actual behavior in the task. If there is a good match, then these reports are amenable to the final step of correlating the outputs of the language models (i.e., the confidence assigned to the different labels) with the overall fit or the parameter of interest in the cognitive model.

Here, we have provided only a conceptual illustration of this technique. In the Appendix, we include a practical example of how to apply this framework to an experiment examining risky choice. This case study primarily demonstrates the rich potential of verbal descriptions for enhancing and verifying cognitive models. As such, it serves as a proof of concept, with certain decisions tailored to our specific task and data set. It also highlights how the failure to elicit sufficiently explicit verbal reports from participants limits the validation process. With improvements in both elicitation techniques and the sophistication of NLP models, we envisage that future attempts to apply our framework will yield even greater insights.

It is important to note that the contents of verbal reports can be used to refine experimental designs and to develop and test new (or alternative) models and theories (see Figure 2: orange node.). The revelation of new, untested models unfolds naturally by employing this method. As participants articulate their thought processes, unforeseen patterns, perspectives, or nuances may emerge. These insights can be pivotal in highlighting gaps in current models or suggesting entirely new avenues of exploration. Moreover, the iterative process of refining experimental designs based on these verbal reports ensures that subsequent studies are better aligned with the intricacies of human cognition and behavior.

Overall, such data can contribute to understanding what might otherwise be considered unexplained noise (Davis-Stober & Regenwetter, 2019; Regenwetter et al., 2022; Regenwetter & Robinson, 2017). Individuals may exhibit variability in behavior because they discovered novel solutions to the problems they are trying to solve or because they shifted their attention to unexplored details in their environment (Szollosi et al., 2023). In other words, the unexplained noise could be the outcome of deliberate, intentional, and well-planned behavior, rather than the "trembling hand" that is often invoked to explain variability.

Discussion

We propose a framework that connects text-based data with the cognitive processes assumed by computational models. The framework assumes that individuals are often aware of their decision strategies, and with the right questions, can provide valuable input about them in form of verbal reports. The responses can be classified by NLP models and can later validate the decision processes the cognitive models assume. In the remainder of the article, we consider two challenges for implementing our conceptual framework.

Challenges in Eliciting Verbal Reports

Perhaps the main challenge in the application of our framework is the need for high quality data, in which people articulate their strategies and reasoning in a way that can be related to cognitive models. Researchers must think about appropriate questions and innovative ways to encourage participants to describe their reasoning in detail.

Overcoming this challenge requires that verbal reports are elicited in a timely fashion and that questions, or elicitation prompts are sensitive to the information participants used to perform a task (e.g., Newell & Shanks, 2014, 2023; Shanks

& John, 1994). Timely collection could imply concurrent elicitation—a method for minimizing information loss from memory or recall biases—but this could extend the duration of the task and prove unsuitable for tasks that prioritize speed and accuracy.

Conversely, retrospective elicitation may lead to reconstructed hypothesized accounts of how a task was approached since the timing at which the verbal reports are elicited allows for additional processing, or post hoc inference. These considerations underscore the essential role the nature of the task plays in determining the appropriateness of the method and its potential effects on both performance and the reliability of verbal reports.

Earlier studies suggest some remedies to these problems (Ranyard & Svenson, 2011; Williamson et al., 2000). For example, Williamson et al. (2000) proposed asking participants to provide verbal descriptions pre- and posttask completion. In the first stage, participants are asked to prepare a list of the information they think they will need to know in order to reach their decision. After the completion of the task (or trial), participants are asked to summarize how they made each of their decisions (Ericsson & Simon, 1980, 1984). A related technique is to ask participants to describe their behavior to a hypothetical future participant in such a way that the future participant will maximize earnings or accuracy in a task.

Other approaches to reduce the measurement errors or imperfections associated with using verbal reports as data are to increase the frequency with which data is collected, to vary prompts or frames for instructions (Gonzalez et al., 2005) and to automate the collection of reports. We developed a tool that addresses these challenges by enabling real-time speech recognition, where spoken reports are automatically converted into text and analyzed using large-language models, greatly reducing the typical burden of analyzing self-report data (Ostrovsky et al., 2024). This last approach offers a practical solution which can assist in synchronizing verbal reports with other behavioral records. This enables the verification of report consistency with both primary behavioral data, and the aspects of the cognitive model used to explain that behavior. Together this can potentially enhance the overall reliability of the information gathered. Such automatic methods can reduce the verbalization-effort experienced by participants and leave cognitive processes engaged by the primary task relatively unaffected (Ericsson & Simon, 1980).

An even more direct method for assessing the accuracy/validity of verbal reports is to compare the performance of separate groups of participants who are directed to behave in different ways on the same task. Consider a simple risk-elicitation task (akin to the one described in our case study in the Appendix) in which participants could be instructed to behave in a "risk-averse" or "risk-seeking" manner. Such a design would enable a comparative analysis of the reports generated under these varied instructions. The hope is that the instructions would lead to identifiable signatures in the behavioral data that would be linked to both elements of the verbal reports and the parameters used to index those behaviors in the computational models. In a sense, this approach "reverse-engineers" the elicitation problem permitting validity and robustness checks across the three sources of input for the analysis.

Challenges in Selecting and Applying NLPs

Just as eliciting accurate verbal reports can be challenging, the selection of the appropriate NLP model for analyzing those reports is highly dependent on the specific nature of the experimental task and the needs of the experimenter. The rapidly increasing progress in the field of NLP models also makes any advice in this area quickly redundant; nonetheless, here, we provide a few pointers to help understand this evolving landscape.

One popular current model is Bidirectional Encoder Representations from Transformers-Multigenre Natural Language Inference is adept at discerning logical connections within verbal communication and offers detailed understanding of complex language nuances (Bhargava et al., 2021). This makes it suitable for analyses requiring deep linguistic insights. Conversely, DistilBERT is noted for its operational efficiency and broad applicability, which is advantageous in diverse experimental settings, especially where there are computational limitations. The choice between these models should be tailored to the unique requirements and constraints of each research project.

The zero-shot classification technique for tasks such as categorizing participants' verbal descriptions stands as the premier choice for classifying new texts without the need for model training specific to a task (Chen & Li, 2021). Therefore, it is well-suited for experimental tasks that produce unique and infrequently encountered data types, as well as for situations where data availability is limited. In the process of classifying participants' verbal reports, as detailed in our case study (see Appendix), we adopted the zero-shot classification technique. This approach was chosen to bypass the need for task-specific model training, a challenge compounded by the scarcity of our data.

A key leverage point of NLP models is in scenarios where numerical data is sparse or unavailable. Word and sentence embeddings present a valuable alternative in these settings as NLP models can convert textual information into a numerical format that computational models can process. An illustration of how NLP models can be leveraged in the context of naturalistic decision making is provided by Bhatia and Stewart (2018). In their experiments, participants were presented with realistic choices between an array of movies and various food dishes, the details of which were meticulously scraped from online sources.

The principal objective of the study was to observe and analyze the cognitive processes employed by participants when confronted with real selections of movie options and culinary dishes. By applying models like those discussed by Bhatia and Stewart (2018), they were able to approximate how individuals evaluate options and make choices in natural contexts. Such models analyze the semantic content of language used by individuals when discussing their preferences, thereby revealing underlying patterns and factors that influence their decisions.

Concluding Remarks

Analyzing verbal reports can arguably be done by human raters. However, even if raters' subjectivity can be overcome by asking them to provide an independent opinion on each text-based response, once verbal description becomes a standard source of data with multiple data points per individual, rating becomes a tedious exercise. This is particularly true when dealing with large data sets or when attempting to analyze the data over time, as the sheer volume of data can quickly overwhelm even the most diligent human rater.

Our novel framework offers significant advantages over these standard approaches in terms of scalability, efficiency, and more objectivity. As such, we believe that automated algorithms for analyzing verbal reports represent an important step forward in learning about how human articulation of strategies can be used to validate, refine, and inspire novel computational models of cognition.

References

Beilock, S. L., & Carr, T. H. (2005). When high-powered people fail: Working memory and "choking under pressure" in math. *Psychological Science*, *16*(2), 101–105. https://doi.org/10.1111/j.0956-7976.2005.00789.x

Bhargava, P., Drozd, A., & Rogers, A. (2021). *Generalization in nli: Ways (not) to go beyond simple heuristics*. arXiv. https://doi.org/10.48550/arXiv.2110.01518

Bhatia, S. (2022). *Inductive reasoning in minds and machines*. PsyArXiv. https://doi.org/10.31234/osf.io/hkpm3

Bhatia, S., & Richie, R. (2022). Transformer networks of human conceptual knowledge. *Psychological Review*, *131*(1), 271–306. https://doi.org/10.1037/rev0000319

Bhatia, S., & Stewart, N. (2018). Naturalistic multi-attribute choice. *Cognition*, *179*, 71–88. https://doi.org/10.1016/j.cognition.2018.05.025

Binz, M., & Schulz, E. (2022). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, *120*(6), Article e2218523120. https://doi.org/10.1073/pnas.2218523120

Binz, M., & Schulz, E. (2023). *Turning large language models into cognitive models*. arXiv. https://doi.org/10.48550/arXiv.2306.03917

Brown, S. D., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*(3), 153–178. https://doi.org/10.1016/j.cogpsych.2007.12.002

Carter, R. M., Meyer, J. R., & Huettel, S. A. (2010). Functional neuroimaging of intertemporal choice models: A review. *Journal of Neuroscience, Psychology, and Economics*, *3*(1), 27–45. https://doi.org/10.1037/a0018046

Chen, C.-Y., & Li, C.-T. (2021). *Zs-bert: Towards zero-shot relation extraction with attribute representation learning*. arXiv. https://doi.org/10.48550/arXiv.2104.04697

Davis-Stober, C. P., & Regenwetter, M. (2019). The "paradox" of converging evidence. *Psychological Review*, *126*(6), 865–879. https://doi.org/10.1037/rev0000156

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional* *transformers for language understanding*. arXiv. https://doi.org/10.48550/arXiv.1810.04805

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

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. MIT Press.

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, *137*(2), 316–344. https://doi.org/10.1037/a0021663

Gluth, S., Kern, N., Kortmann, M., & Vitali, C. L. (2020). Value-based attention but not divisive normalization influences decisions with multiple alternatives. *Nature Human Behaviour*, *4*(6), 634–645. https://doi.org/10.1038/s41562-020-0822-0

Gonzalez, C., Dana, J., Koshino, H., & Just, M. (2005). The framing effect and risky decisions: Examining cognitive functions with fmri. *Journal of Economic Psychology*, *26*(1), 1–20. https://doi.org/10.1016/j.joep.2004.08.004

Krajbich, I. (2019). Accounting for attention in sequential sampling models of decision making. *Current Opinion in Psychology*, *29*, 6–11. https://doi.org/10.1016/j.copsyc.2018.10.008

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, *108*(33), 13852–13857. https://doi.org/10.1073/pnas.1101328108

Lagnado, D. A., Newell, B. R., Kahan, S., & Shanks, D. R. (2006). Insight and strategy in multiple-cue learning. *Journal of Experimental Psychology: General*, *135*(2), 162–183. https://doi.org/10.1037/0096-3445.135.2.162

Navarro, D. J. (2021). If mathematical psychology did not exist we might need to invent it: A comment on theory building in psychology. *Perspectives on Psychological Science*, *16*(4), 707–716. https://doi.org/10.1177/1745691620974769

Newell, B. R., & Shanks, D. R. (2014). Unconscious influences on decision making: A critical review. *Behavioral and Brain Sciences*, *37*(1), 1–19. https://doi.org/10.1017/S0140525X12003214

Newell, B. R., & Shanks, D. R. (2023). *Open minded: Searching for truth about the unconscious mind*. MIT Press.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231–259. https://doi.org/10.1037/0033-295X.84.3.231

Ostrovsky, T., Liew, S. X., & Newell, B. R. (2023). *How do people integrate private and social information when making risky decisions?* [Manuscript in preparation]. University of New South Wales.

Ostrovsky, T., Ungermann, P., Newell, B. R., & Donkin, C. (2024, March). *From verbal reports to model validation: Theoretical framework and application* [Conference session]. Conference of experimental psychologists, Regensbury, Germany.

Proulx, T., & Morey, R. D. (2021). Beyond statistical ritual: Theory in psychological science. *Perspectives on Psychological Science*, *16*(4), 671–681. https://doi.org/10.1177/17456916211017098

Ranyard, R., & Svenson, O. (2011). Verbal data and decision process analysis. In M. Schultze-Mecklenbeck, A. Küberger, & J. G. Johnson (Eds.), *A handbook of process tracing methods* (2nd ed., pp. 270–285). Routledge.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108. https://doi.org/10.1037/0033-295X.85.2.59

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

Regenwetter, M., & Robinson, M. M. (2017). The construct–behavior gap in behavioral decision research: A challenge beyond replicability. *Psychological Review*, *124*(5), 533–550. https://doi.org/10.1037/rev0000067

Regenwetter, M., Robinson, M. M., & Wang, C. (2022). Four internal inconsistencies in tversky and kahneman's (1992) cumulative prospect theory article: A case study in ambiguous theoretical scope and ambiguous parsimony. *Advances in Methods and Practices in Psychological Science*, *5*(1). https://doi.org/10.1177/25152459221074653

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter*. arXiv. https://doi.org/10.48550/arXiv.2110.08207

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Le Scao, T., Raja, A., Dey, A., Saiful Bari, M., Xu, C., Thakker, U., Sharma, S., Szczchia, E., Kim, T., Chhablani, G., Nayak, N., … Rush, A. M. (2022). *Multitask prompted training enables zero-shot task generalization*. arXiv. https://doi.org/10.48550/arXiv.2110.08207

Shanks, D. R., & John, M. F. S. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences*, *17*(3), 367–395. https://doi.org/10.1017/S0140525X00035032

Singmann, H., Kellen, D., Cox, G. E., Chandramouli, S. H., Davis-Stober, C. P., Dunn, J. C., Gronau, Q. F., Kalish, M. L., McMullin, S. D., Navarro, D. J., & Shiffrin, R. M. (2022). Statistics in the service of science: Don't let the tail wag the dog. *Computational Brain & Behavior*, *6*, 64–83. https://doi.org/10.1007/s42113-022-00129-2

Svenson, O. (1979). Process descriptions of decision making. *Organizational Behavior and Human*

Performance, *23*(1), 86–112. https://doi.org/10.1016/0030-5073(79)90048-5

Svenson, O. (1989). Eliciting and analyzing verbal protocols in process studies of judgment and decision making. In H. Montgomery & O. Svenson (Eds.), *Process and structure in human decision making* (pp. 65–81). Wiley.

Szollosi, A., Donkin, C., & Newell, B. R. (2023). Toward nonprobabilistic explanations of learning and decision-making. *Psychological Review*, *130*(2), 546–568. https://doi.org/10.1037/rev0000355

Szollosi, A., & Newell, B. R. (2020). People as intuitive scientists: Reconsidering statistical explanations of decision making. *Trends in Cognitive Sciences*, *24*(12), 1008–1018. https://doi.org/10.1016/j.tics.2020.09.005

Vanunu, Y., Hotaling, J. M., Le Pelley, M. E., & Newell, B. R. (2021). How top-down and bottom-up attention modulate risky choice. *Proceedings of the National Academy of Sciences*, *118*(39), Article e2025646118. https://doi.org/10.1073/pnas.2025646118

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

Williamson, J., Ranyard, R., & Cuthbert, L. (2000). A conversation-based process tracing method for use with naturalistic decisions: An evaluation study. *British Journal of Psychology*, *91*(2), 203–221. https://doi.org/10.1348/000712600161790

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davidson, J., Shleifer, S., van Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. M. (2019). *Huggingface's transformers: State-of-the-art natural language processing*. arXiv. https://doi.org/10.48550/arXiv.1910.03771

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davidson, J., Shleifer, S., van Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., … Rush, A. M. (2020). *Transformers: State-of-the-art natural language processing* [Conference session]. Empirical methods in natural language processing: System demonstrations. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Yin, W., Hay, J., & Roth, D. (2019). *Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach*. arXiv. https://doi.org/10.48550/arXiv.1909.00161

Appendix

Case Study Details

This Appendix details our case study (which serves as an illustration of our theoretical framework). It includes the description of our experimental design, our behavioral data, the procedure of exclusion criteria applied in our analysis, the fitting of the *Bayesian updating* and the *social distance* models and the information about the NLP model (and pipeline) used to analyze the verbal results, the candidate keywords used as the input for the NLP model as well as the results of this analysis. It also contains the raw verbal descriptions participants provided across two experiments.

A Case Study

Before describing our case study in detail, we note that we anticipate a significant portion of the methodological steps we outline here might not be required once NLP models have been trained and fine-tuned to categorize verbal descriptions of participants' strategies. In other words, the primary goal of this case study is to show that verbal descriptions represent a valuable reservoir of data, which can be both enlightening and beneficial for verifying cognitive models.

Our proposed technique was assessed based on existing data comprising verbal descriptions of choice strategies provided by participants in two experiments conducted as part of the first author's PhD dissertation. In essence, the current analysis attempts to map these descriptions onto the psychological assumptions made by one of the two computational models—the *distance model*—that we proposed and tested in that work (Ostrovsky et al., 2023, article in preparation). We describe the methods and results of this work below. In the full project, we also considered an additional *Bayesian updating model* that incorporated different underlying assumptions. For brevity, we focus here on the validation of the *distance model* but all details about the application of our technique to the *Bayesian updating model* can be found in the Appendix.

(*Appendix continues*)

Methods and Procedure

Behavioral Data

As an initial step, we evaluate the appropriateness of the experimental task for eliciting verbal reports, as outlined in Figure 1. Our task was not subject to time constraints and appears to facilitate the accessibility of information regarding the strategies used by individuals. In the subsequent section, we detail the task undertaken by participants to illustrate the aspects we anticipated would be verbalizable within this task.

Participants were presented with the same risky-choice problem in every trial: There were two jars: Jar A consisted of 300 black balls, each worth £14 and £700 white balls, each worth £0. Jar B consisted of a flipped distribution of balls with 700 black balls, each worth £6 and £300 white balls, each worth £0. Hence, the two jars had equal expected value (henceforth expected value) but differed in their riskiness.[A1] Participants were told that one of two jars will be randomly selected by the computer on each trial. As participants were not told which of the two was selected, they were presented with a 100-ball sample (from the selected jar) to help them work out which jar was selected. They were then asked to choose from which jar they wish to draw their outcome-determining ball. They could either choose to select from the one indicated by the contents of the 100-ball sample (by pressing the button “stick”) or to choose to draw from the unselected jar (by pressing the button “switch”).

Choosing to draw from Jar A indicates an appetite for risk and choosing Jar B indicates an aversion to risk. Risk seeking individuals are those, who chose Jar A (either by choosing to stick to this jar or by actively switching to draw from it) on the majority of trials. Risk averse individuals are those, who chose Jar B (either by choosing to stick to this jar or by actively switching to draw from it) on the majority of trials.

Both experiments used a within-subjects design with two conditions (*solo* and *social*) each presented in their respective block of trials. In the first block, participants made a series of risky choices alone (i.e., *solo* condition; see Figure A1 panel A). In the second block, participants completed a series of the same risky decisions as in the *solo* condition, however, this time they made their decisions after observing choices made by other (hypothetical) people (i.e., *social* condition; see Figure A1 Panel B). Trials differed in terms of the composition of the 100-ball sample shown to participants, and the number of other players choosing to take or avoid risks. For brevity, here we focus on how participants’ choices changed between the solo and social phases of the experiment.

Modeling Data

As our second step, detailed in Figure 1, we begin this section with the delineation of the models that were defined and fitted to the data. Furthermore, we explicate the psychological interpretation of their parameters, setting the stage for the subsequent analysis of the verbal reports.

To explain how participants integrate and update their private information (i.e., in the form of the 100-ball sample) given social information (i.e., the choices made by the hypothetical others), we analyzed two primary models. The model that consistently outperformed in various experiments is referred to as the *distance model* which posits that people have a sense of how different they are from others and they update their choices based on how much weight they put on the distance between themselves and others.

$$P(C) = \left[\beta \times P\left(S - \frac{n}{N}\right)\right] + P(S). \qquad \text{(A1)}$$

Equation A1 describes the *Distance* model. $\frac{n}{N}$ presents the proportion of social evidence. n are the number of choices others made in favor of one of the choice options and N is the total number of choices made. $P(S)$ is proportion of choices for that option made in the absence of social information. β is a free parameter representing the weight people assign to the *relative* distance between their own choices and the choices made by others (i.e., social information). It has a lower boundary of $\beta = 0$ representing ignoring others and accounting only for one’s own preferences and an upper boundary $\beta = 1$ indicating a full integration of others’ choices into one’s own.

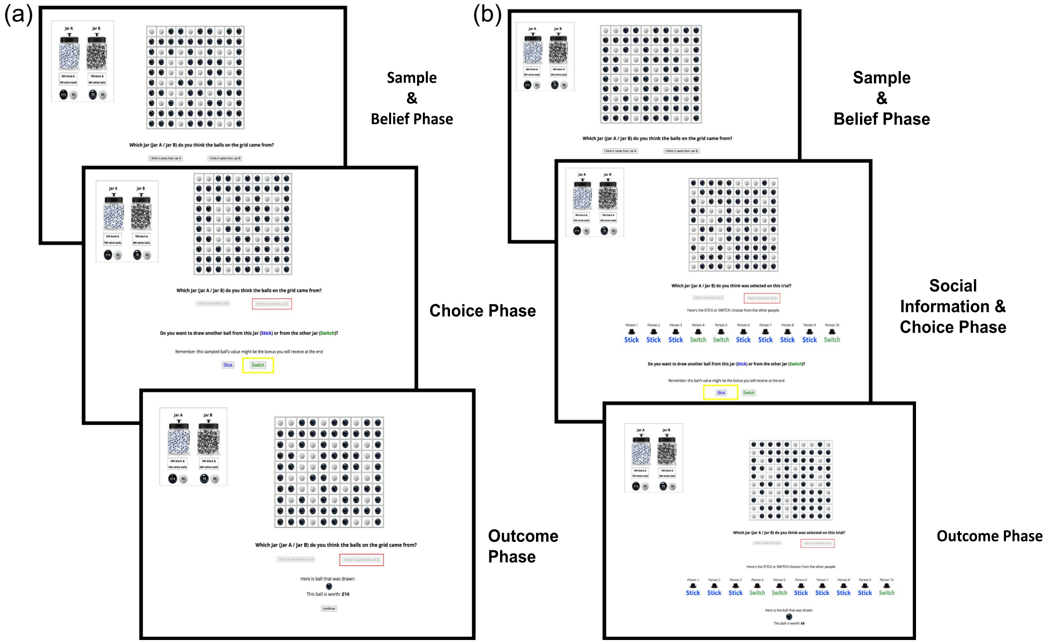
In a nutshell, our results can be summarized as the following: (a) people pay attention to others’ choices (to different degrees) and (b) information integration is consistently inline with the *relative updating* principle. That is, the *distance* model

[A1] $\text{EV}_{\text{JarA}} = \text{EV}_{\text{JarB}} = 4.2$; $\text{Variance}_{\text{JarA}} = 41.16$, $\text{Variance}_{\text{JarB}} = 7.56$.

(*Appendix continues*)

Figure A1

Trial Layout for the Experiments


(a)
Jar A
Jar B
300 black & 700 white balls
700 black & 300 white balls
Which Jar (Jar A / Jar B) do you think the balls on the grid came from?
Do you want to draw another ball from this Jar (Stick) or from the other Jar (Switch)?
Remember: this sampled ball's value might be the bonus you will receive at the end
Stick
Switch
Here is ball that was drawn:
This ball is worth: £14
continue
Sample & Belief Phase
Choice Phase
Outcome Phase
(b)
Which Jar (Jar A / Jar B) do you think was selected on this trial?
Here's the STICK or SWITCH choices from the other people:
Person 1
Person 2
Person 3
Person 4
Person 5
Person 6
Person 7
Person 8
Person 9
Person 10
Stick
Stick
Stick
Switch
Switch
Stick
Stick
Stick
Stick
Switch
Do you want to draw another ball from this Jar (Stick) or from the other Jar (Switch)?
Remember: this ball's value might be the bonus you will receive at the end
Stick
Switch
Sample & Belief Phase
Social Information & Choice Phase
Outcome Phase

Note. Panel A shows steps from the solo condition in which participants indicated their risk preferences (choice of a jar to draw from) in the absence of any information about what other players chose. Panel B shows steps from the social condition in which participants indicated their risk preferences after seeing what 10 hypothetical other players had chosen to do. See the online article for the color version of this figure.

provided a better account of the data for the majority of participants than the alternative model (the Bayesian updating model), details of which can be found in the Appendix.

Crucially, at the end of the experiment, participants were asked to describe any choice strategies they used in the first block (i.e., when making risky decisions alone) and those they used in the second block (i.e., when making decisions after observing the choices made by the other hypothetical players). Participants inserted their verbal descriptions into an empty text box that was not restricted by a maximum length. It is noteworthy that the verbal descriptions elicited at the end of the two experiments were *not* collected for the purpose of this NLP analyses described here but were rather included, as in many psychological experiments, as additional source of information that *could* be of use.

There were two main consequences of this design decision. First, since our questions were not specified for this task, not all participants provided descriptions that could be classified as a choice description. For instance, consider the description given by one participant about a strategy used in the *solo* condition: "I relied on how many black or white balls it looked like there were." This individual is clearly describing *a* strategy. However, this description does not allow for any inference about their intentions to chose Jar A or Jar B. It does not, for example, specify which color of balls (i.e., black or white), or if any at all, determined their choice of jar. Unfortunately, in these occasions, there was no meaningful label that could be matched to a choice behavior (i.e., to the number of risky choices) and we had to exclude those relatively early in our analysis procedure. Second, as our participants were not directly asked to, they rarely differentiated between a strategy description (i.e., a *how* question) and their reasoning (i.e., a *why* question). In the Appendix, we provide a full summary of the verbal descriptions as well as the best suitable label assigned for each individual (see Appendix, Tables A7, A8 and A9).

(*Appendix continues*)

Despite these challenges, our experimental task and analysis provide three important sources of information: (1) behavioral data about participants' preferences for risk when making decisions alone (from the *solo* condition) and after observing others' choices (from the *social* condition), (2) individuals' model fits, and (3) text-based descriptions of strategies provided by our participants. All sources of information are crucial for the mapping between NLP model output and our individual model fits, making the current data set a suitable, though by no means perfect, "toy" example for our framework.

Current NLP Model

In this section, we explain our methodological steps for this case study. As mentioned above, some of these steps are unique for this case study and are outlined in detail to allow the reader to follow our work flow. Later we discuss which steps should become unnecessary once NLP models are trained to learn to classify verbal descriptions. Figure 1 illustrates, once again, how our case study serves as an exemplar for the application of this framework. Given that our analysis was conducted retrospectively, the considerations regarding the elicitation point, reporting structuring type, the formulation of instructions, and the cadence of elicitation have been predetermined. However, as highlighted in the lilac right rectangle of the figure, we advise that these critical questions be addressed in advance when determining the specific task and the precise methods for eliciting verbal reports.

We used a modified version of the bidirectional encoder representations model (henceforth: BERT); a new type of neural network transformer model for natural language processing that was pretrained on billions of words (Devlin et al., 2018). For our specific purpose of labeling, we used the *DistilBERT* model, which is a compacted and quicker version of the full BERT model. Using this version, increases the efficiency and reduces the compute-time needed to run these types of models (Sanh et al., 2019).

As our data set was not large enough to fine tune the BERT model from scratch we used a preexisting fine-tuned version of BERT provided by the *Huggingface* company, which is a curated collection of pretrained models (Wolf et al., 2019). The fine-tuned model we selected for our labeling task is called "zero-shot-classification" pipeline. This pipeline allows classification of novel text that was *not* present in the training data and the association of a label (or a set of candidate labels) with a piece of text, irrespective of its domain or its aspect (e.g., emotion, name, event, etc.). In practice, the model takes a sentence and a list of candidate labels as input and returns accuracy score(s) to each tested label(s) (Sanh et al., 2022; Yin et al., 2019). This "accuracy" score ranges between 0 and 1, which can be understood as the "confidence" the model assigns to the ability of a label to describe the sentence (see Figure A2).

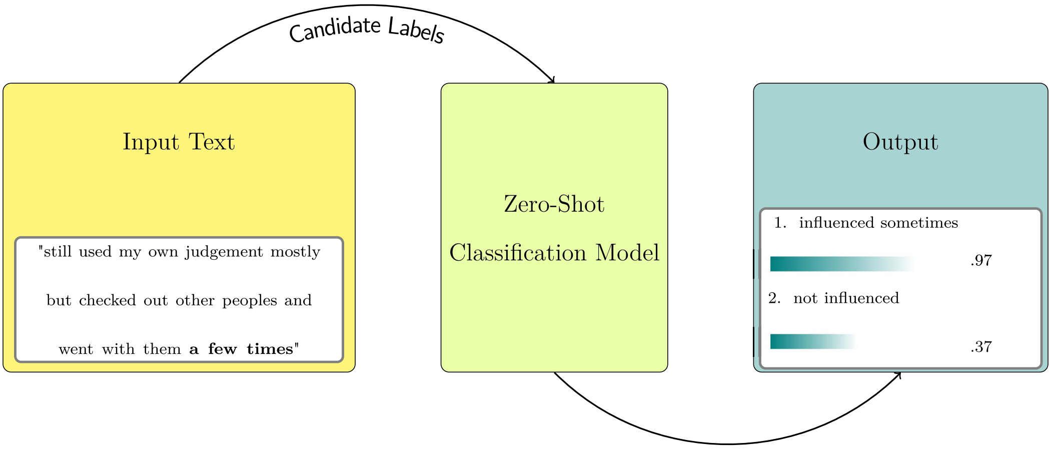
In zero-shot classification, the input text is encoded using a pretrained BERT model, which tokenizes and generates contextual embeddings from the text. Each label category is also transformed into an embedding vector using techniques like averaging word embeddings or the universal sentence encoder. The process then involves computing similarity scores between the text's embedding and the label embeddings, using metrics such as cosine similarity or Euclidean distance. These scores indicate how well the input aligns with each label's context. The label with the highest similarity score is chosen as the input text's predicted class, with each label receiving a score representing this alignment (Chen & Li, 2021).

Labeling—Solo and Social Condition

To classify participants' verbal descriptions into "description strategy" groups, BERT tests the relation between a set of candidate labels and selects the one that best describes this description. Our main goal for this analysis was to correlate these label scores with the individual model fits (that describe the way people integrate information) to assess the validity, or psychological plausibility of the models' core assumptions. To do that, we first needed to identify the quality of the verbal descriptions. We did so by comparing the selected (i.e., most suitable) label of individuals' strategy to their actual behavior in the task (For a complete list of the labels used in the *social* and *solo* conditions, see Tables A3 and A4). For example, if a person's best label on their verbal description indicates that they followed others when making their decisions in the second experimental block (see Figure A1b), we could check if they, indeed, changed their choices, on the majority of the trials, in the *social*

(*Appendix continues*)

Figure A2
A Graphical Representation of the Process, in Which BERT Classifies Text Using "Zero-Shot Classification" Pipeline With an Illustrative Example


Candidate Labels
Input Text
"still used my own judgement mostly but checked out other peoples and went with them **a few times**"
Zero-Shot Classification Model
Output
1. influenced sometimes
.97
2. not influenced
.37

Note. The left yellow box represents the first step, in which text is fed into the model alongside a set of candidate labels. In our example, the model is given the text "still used my own judgment mostly but checked out other peoples and went with them a few times" with candidate labels "influenced sometime" and "not influenced." The middle box represents the fine-tuned model that processes the association between every label with the input text. The teal-colored box on the right represents the output stage, in which end-users obtain the scores the model assigned to each, from best to worst. In this example, the model assigned a high accuracy score to the label "influenced sometimes" and low accuracy to the label "not influenced." BERT = Bidirectional Encoder Representations from Transformers. See the online article for the color version of this figure.

condition relative to the trials in the *solo* condition (see Figure A1a).

Strategy Classification

For the first step in the current analysis, we used the NLP model to classify strategies described in each experimental condition. As our model could not be trained (and fine-tuned) on data to classify decision making strategies, we defined the labels using terms that describe choice strategies in these conditions. The labels were allocated into a set of "strategy-description" classes.

Strategy Classification Versus Actual Behavior

The second step was to compare the way in which participants described their behavior with what they actually did. We conducted this comparison in three steps. First, we selected the labels that the model assigned the highest confidence to (i.e., had the highest score; see Tables A5 and A6, Tables A7, and A8). In a second step, we applied an exclusion criteria to exclude participants, who clearly described *a* strategy but failed to describe it in a way that revealed their choice/preference for one of the scoring strategies. Last, to identify "good" describers, we matched the labels' scores with actual behavior.

Strategy Classification Versus Model Fits

Following these steps, we then matched the weighting parameter (i.e., β) to the NLP labels' scores. To do so, we rank-ordered the NLP model's score of the strategies in the *social* condition in the following way: If the individuals' best label indicated that they did not rely on others' choices (e.g., "did not influence me"), the score should be low when matched with the weight-on-distance parameter. This is because a low β estimates indicates not relying on the choices of others. Therefore, as scores on labels were around .9 (indicating a high confidence that the label described the text well), we subtracted the score from 1 when compared with the weighting parameter (i.e., they put low weight on others). If the individuals' best label indicated that they incorporated others' choices into their own, we left them as is when comparing with the weight-on-distance parameter (i.e., they put high weight on others and thus have a high β estimate). We also coded our labels' scores to individuals

(*Appendix continues*)

who suggested that they were occasionally influenced by others' choices by converting a high score into a mid-weight by subtracting .5 from it. If the labels indicated that others had a frequent influence on choice, we ranked the score to align with a relatively high weight by subtracting .2 from their label accuracy score (i.e., interpreting "often" as .8 of the time).

Note that we asked participants only once about their strategy (at the end of the experiment) and hence relied on a single description per participant. Behavior, on the other hand, was based on the proportion of choice across 60 trials. To address this issue, we allowed up to 30% discrepancy between the labels' scores for the verbal descriptions and the observed behavior. That is, if the distance between actual behavior and the confidence in the label's score was less than .3, participants were included for the analysis. If it was greater than .3, we excluded participants. The full description of the exclusion criteria are summarized in the Appendix.

In applying our discrepancy criteria to verbal reports from both solo and social settings, we observed a substantial decrease of the data of around 25%, which is less than ideal. This reduction might be attributed to the fact that our analysis was conducted post hoc and reflects that our study was not initially designed to focus on verbal reports. Adhering more closely to Ericsson and Simon's (1980) methodology from the outset might have mitigated this loss of data.

Tables A5 and A6 and Figure A5 summarize these steps both graphically and descriptively. We discuss how to minimize this difference in measures of behavior for future projects in the Discussion section.

Results

We used three sources of data: empirical data, individual model fits, and the verbal descriptions on decision strategies provided by participants. We report the results on the link between the NLP outputs and the *distance* model outputs. The results of the same analysis using the *Bayesian updating* model output are provided in the below (see, Table A2 and Figure A4).

Verbal Descriptions and the Distance Model

We analyzed whether individual fits of the *distance* model (see Equation A1) are validated by the descriptions individuals provided about their choice strategies. Recall that this model assumes that information is integrated in *relative* terms. The free parameter β represents the weight on the distance between risk preferences exhibited alone and the proportion of other hypothetical players who chose the risky option in the social phase. It has a lower boundary of $\beta = 0$ representing ignoring others and paying attention only to one's own preferences and an upper boundary $\beta = 1$ representing full integration of others' choices into one's own. Figure A3 displays the individual fits of the *distance* model as a function of the NLP model classification of their verbal descriptions. Figure A3a shows the results of the full data set consisting of all individuals' verbal reports and their model fits. Figures A3b and A3c displays the results from the data after applying the exclusion criteria (see Methods).

These figures show three key patterns. First, we observe a greater density around the values 0 and .5 on the *x* axis relative to the values around 1.0. This can be explained by participants' conservative descriptions about the reliance on others. They do not often describe their choice as completely affected by others' behavior. However, note that those individuals were removed after applying our exclusion criterion (i.e., applied to those who did not provide a verbal description that was in line with their behavior in both the solo and social phases of the experiment).

We tested our visual impression with a linear model for each data set separately to learn if the verbal descriptions, were, indeed predictive of the psychological assumptions in the *distance* model (i.e., that people integrate others choices depending on the weighted distance between others and oneself). Our results are summarized in Table A1 and suggest that the scores of the verbal descriptions are a significant predictor of the weighting *distance* model parameter only after applying our exclusion criteria. Notably, this leaves us with only a quarter of our sample, an issue that we return to in the Discussion.

Exclusion Criteria

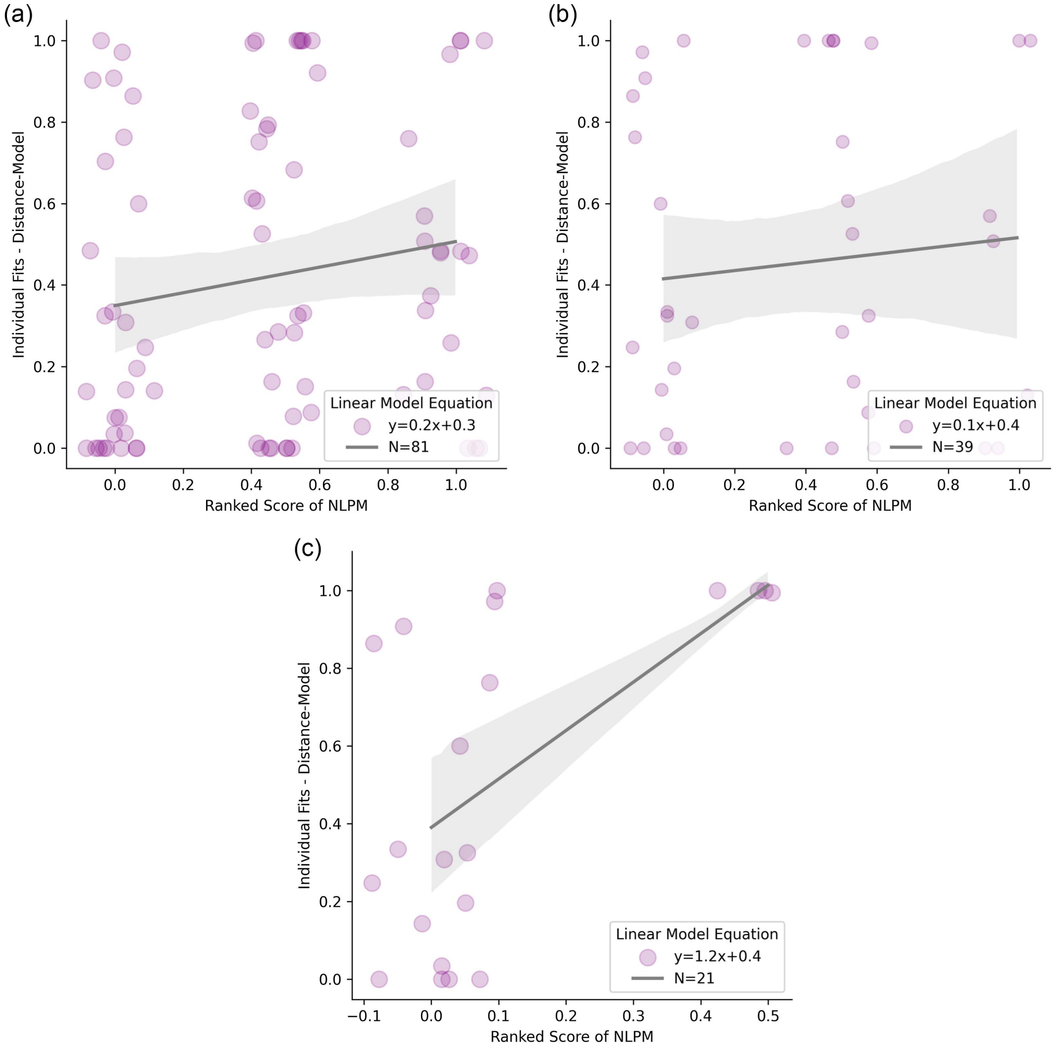
Exclusion Criterion 1

Labeling the text participants generated allows us to compare the way in which they described their behavior with what they actually did. We conducted this comparison in three steps. First,

(*Appendix continues*)

Figure A3

The Relationship Between NLP Model's Output (Rank-Ordered Labels' Scores) and the Distance Model's Individuals Fits



Note. (a) The relationship computed based on a full data set (no excluded describers). (b) The relationship computed based on a subset of the full data set (excluded describers based on the quality of description in the solo condition). (c) Excluded describers based on the quality of description in the solo condition and social condition. Note that the *distance* model's free parameter is the weighing parameter β with 0 and 1 for the lower and upper boundaries, respectively. Note that higher values of β (weighting in *distance* model) indicate greater influence of social information, whereas higher values of *r* (social resistance in *Bayesian updating* model) indicate lesser influence of social information. NLP = natural language processing; NLPM = natural language processing model. See the online article for the color version of this figure.

we selected the labels that the model assigned the highest confidence to (i.e., had the highest score; see Tables A1 and A2). In a second step, we applied an exclusion criteria to exclude participants, who clearly described *a* strategy but failed to describe it such that it revealed their choice/preference for one of three strategies: (1) a preference for risk aversion, (2) a preference for risk seeking, or (3) no clear preference (i.e., a description of a rather random strategy such as "I went with my gut feeling"). To identify "good" describers we matched the labels' scores with

(*Appendix continues*)

Table A1
Regression Model

Data set	Intercept	β estimate	*p*	R^2	*N*
Full	.37	.16	.157	.025	81
Exclusion Criterion 1	.33	.09	.560	.009	39
Exclusion Criterion 2	−.05	.28	**.005**	.345	21

Note. Value presented in bold indicate statistical significance, with an alpha level of less than .05. Text analysis and the cognitive model's individual fits. Regression estimates from the text analysis and distance model individual fits. *N* refers to the number of individuals included in the analysis.

actual behavior. That is, a high confidence in a preference for risk seeking (risk aversion) was compared to the number of risky (risk averse) choices individuals made during the *solo* condition. For example, the description: "I tried to maximize my chance to get the highest award" was classified as risk seeking since the highest reward could have been obtained by choosing the riskier jar (i.e., choosing to draw from jar A). If the participant indeed chose the riskier option in the majority of the trials (i.e., in more than 50%), she was identified as a "good (risk seeking) describer." For those, who were best described by a "random" labels, we matched with a rather indifference to risk (i.e., about 50% risky choices). Clearly, these nonclassifiable participants perhaps have tried to describe their behavior but failed to do so in a way that captured their behavioral patterns. This is likely a result of the way, in which we formulated our questions. We return to address this issue in the Discussion section below. We acknowledge that verbal descriptions, just as behavior, can be noisy. We, however, asked participants only once and hence relied on a single description. Behavior, on the other hand, was based on the proportion of choice across many trials. To address this issue, we allowed up to 30% discrepancy between verbal descriptions labels scores and behavior. That is, if the distance between actual behavior and the confidence in the label's score was less than .3, participants were included for the analysis and excluded otherwise. Tables A1 and A2 and Figure A5 summarize these steps both graphically and descriptively. We discuss how to minimize this error in future projects in the Discussion section.

Exclusion Criterion 2

As in the *solo* condition, we defined "good describers" as participants whose descriptions could be validated by their actual behavioral data. This validation was computed as the (absolute) difference between the risk preferences revealed in the *solo* condition and those revealed in the *social* condition. Finally, as before, we allowed some error to exist between verbal descriptions and actual behavior (i.e., error margin of up to 30%). If participants' reports were consistent with a change in behavior between the *solo* condition *and* in the *social* condition that was within the determined error margin,[A2] they were classified as "good describers" and were included in the last step for this analysis and were excluded otherwise. To summarize, in the second and last exclusion step in the current analysis, only "good describers" identified from *both* the *solo* and the *social* condition (see results, Figures A3c and A4c) were included.

We further broke down each main class into subclasses, that represented a specific task feature that was used to describe a decision process. Recall that participants could use many cues to make their decisions. They could rely on the distribution of the balls, on their outcomes or a combination of the two. We first added a subclass that simply described an explicit expression of a preference for a specific jar (e.g., "chose Jar A"/"chose Jar B"). We then added terms that were related to choices based on the distribution of the balls (e.g., "more black balls," "jar with the most white"), and those based on outcomes (i.e., "get 6," "get higher value"). For random/indifference descriptions, we included labels such as "gut feeling" or "intuition." For example, while some people may focus on their chance to receive a/any reward, others might focus on the chance to obtain a specific reward. Two illustrative examples are: "I chose the jar with the majority of black balls" or by another

[A2] This computation was on absolute scale.

(*Appendix continues*)

Table A2

Regression Estimates From the Text Analysis and Bayesian Updating Model Individual Fits

Data set	Intercept	r estimate	p	R^2	N
Full	.54	−.02	**.04**	.053	81
Exclusion Criterion 1	.46	−.02	.19	.046	39
Exclusion Criterion 2	.25	−.03	**.002**	0.401	21

Note. Values presented in bold indicate statistical significance, with an alpha level of less than .05. *Data set* refers to the data set applied that generated the respective subfigure. *r Estimate* refers to the estimated slope examining the relation between the ranked-score of language label and the r parameter in the Bayesian model. R^2 refers to the percentage of the variance in the r parameter that is explained by the language label score. N refers to the number of individuals included in the analysis.

participants as: “I tried to maximize my chance to win.” These descriptions suggest that these people used different sources of information to indicate a preference for the same jar (Jar B). That is while the first description focused on the distribution of the (black) balls, the second focused on the chance to win a reward (which is maximized in Jar B). In contrast, we obtained other descriptions in line with a preference for the riskier Jar A; examples include: “I chose the jar with the majority of white balls” and “I tried to maximize my chance to get the highest award.” Table A4 contains the full list of all the main and subclasses of the labels we tested in this analysis.

Bayesian Model

$$P(C) \propto P(S) \times \left[\int_{.5}^{1} B(n, N, p)dp + r\right]. \quad \text{(A2)}$$

Equation A2 describes the *Bayesian updating* model. $P(C)$ is the posterior probability of choosing the riskier option in the social information condition. $P(S)$ is the prior preference (i.e., when choosing without exposure to any social information) for one of the choice options. $B(.)$ is the binomial probability mass function for observing n_j successes or pieces of evidence out of a total of N.[A3] r is a freely varying resistance parameter that measures the extent to which people inhibit the updating of their prior preference. If $r = 0$, no resistance is present and the individual maximally updates their behavior based on social information. However, if $r > 0$, the individual is increasingly resistant toward updating such that their posterior choices are minimally affected by social information.

Verbal Descriptions and the Bayesian Updating Model

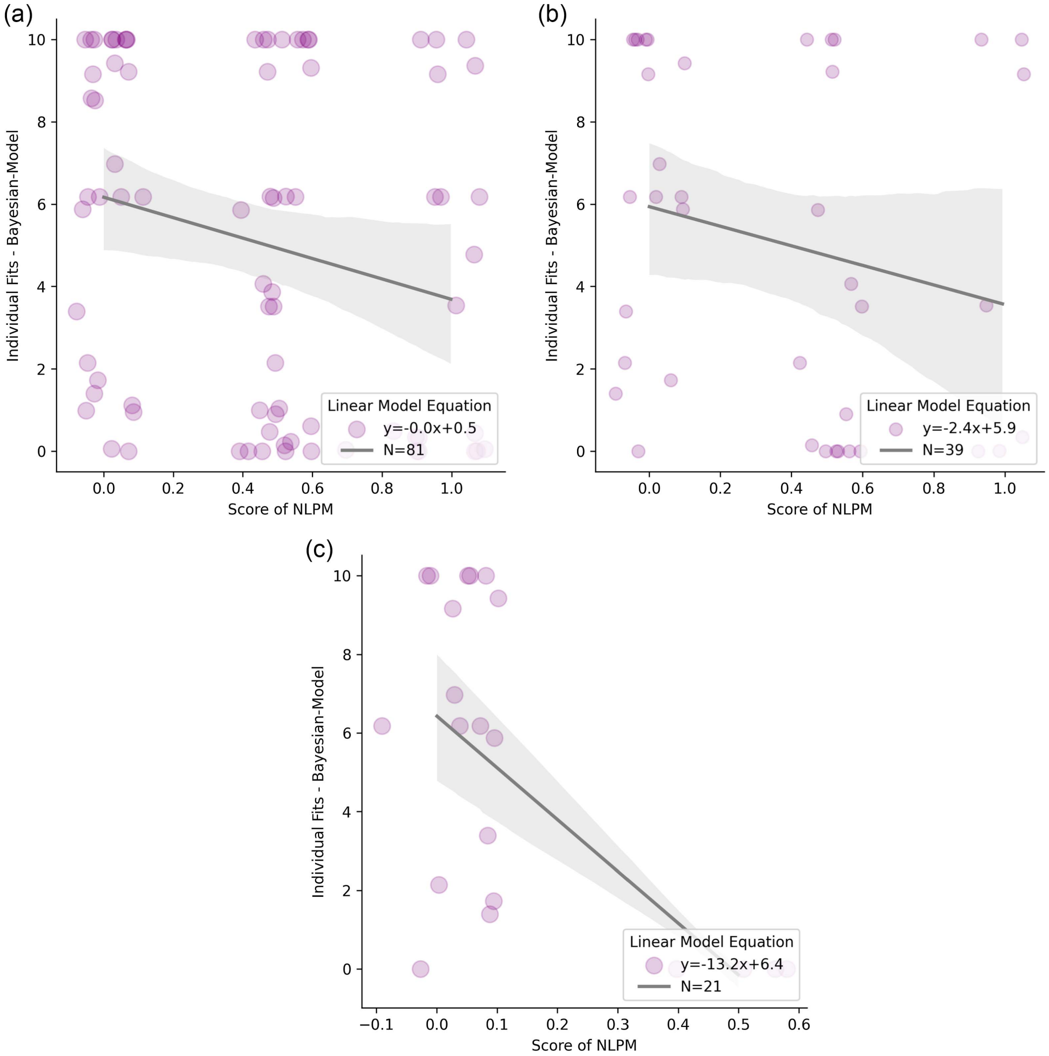
We analyzed whether individual fits of the *Bayesian updating* model (see Equation A2) are validated by individuals’ descriptions of their decision strategies. Recall that this model assumes that information is integrated on an absolute scale. The free parameter r represents the resistance toward updating social information with lower boundary of $r = 0$ representing no resistance (and choices that are fully explained by the observation of others’ choices) and upper boundary $r = 10$ representing high resistance (and choices that are explained fully by one’s own preferences for risk). In our original analysis of the data, this model was outperformed by the *distance* model both on the group and on the individual level. Interestingly, we find similar results to those observed in the *distance* model. Although Figure A4a seems almost identical to Figure A3a, the statistical analysis provides evidence that there is a (weak) correlation between the labels’ scores and the individual fits from the *Bayesian updating* model. This was also true when we test this relationship on the data that consists of “good” describers after the exclusion criteria was applied to the data (see Figure A4c). The results of the linear models are summarized in Table A1. The findings suggest that the language model output is a significant predictor in the full data set and after applying the exclusion criterion.

[A3] This model takes as its likelihood the probability of observing n_j pieces of information out of a total of N.

(*Appendix continues*)

Figure A4

The Relationship Between Language Model's Output and the Bayesian Updating Model Individuals Fits



Note. (a) The relationship computed based on a full data set (no excluded describers). (b) The relationship computed based on a subset of the full data set (excluded describers based on the quality of description in the solo condition). (c) Excluded describers based on the quality of description in the solo condition and social condition. Note that higher values of β (weighting in *distance* model) indicate greater influence of social information, whereas higher values of *r* (social resistance in *Bayesian updating* model) indicate lesser influence of social information. NLPM = natural language processing model. See the online article for the color version of this figure.

Discussion

While our case study data does not provide the ideal test-bed for these ideas, we think they provide a useful proof of concept for the application of NLP models in other tasks.

We observe a consistent relationship between models estimates' and verbal descriptions only when the relatively small subsample of "good" describers are included in the analyses (see Figure A5 and Table A1). These findings indicate that mapping verbal descriptions onto either empirical behavior or the psychological assumptions made by our models is not a straightforward task.

There are two ways to interpret our mixed results. The first takes our results as a challenge to the underlying psychological process assumed by

(*Appendix continues*)

our model. That is, people do not assign a mental weight on the distance between themselves and others but rather act according to a model that was not yet considered. As can be seen in the Appendix, attempts to fit the Bayesian model suggest that people do not reliably follow the absolute updating principle either. This is a valid interpretation and our framework provides suggestions to next possible steps (see Figure 2). One possible step is to use individuals' descriptions (especially those, who included their reasoning) as a guide in thinking about future models or possible modifications to the experimental design. For example, in the present case study, some individuals wrote that after trying to follow the choices made by the hypothetical other participants, they experienced losses or gains indicating that they were trying to learn about (spurious) dependencies between features of the experiment.

However, there is a valid reason to assume that our inconclusive findings are a direct consequence of the formulation of our questions. Judging by the results of the NLP models' text classification, we could see that although many of our participants indeed describe their choice processes, their responses did not contain the necessary information to align exactly with our models' parameters. That is, we did not ask questions that could provide sufficiently rich descriptions. First, the verbal descriptions frequently described behavior alone but did not disambiguate the reasoning behind it. Just like any other scientific theory, a model's aim is to provide not only a description of some (behavioral) regularities but also an explanation for them (Navarro, 2021; Proulx & Morey, 2021; Singmann et al., 2022). In order to test the validity of our cognitive models we would need a description of both (i.e., the description of behavior and the reasons underlying it). Despite our inconclusive results, we believe that posing appropriate questions to our participants that are defined with the theory and models in mind and that include both the question of *how* and the question of *why* will be able to bring new insights.

Labels Used for the NLP Model's Analysis

Table A3
Keywords Used in Social Condition

Meta class	Always	Sometimes	Often	Never
Influence	“influenced by”	“sometimes influenced me”	“often influenced me”	“did not influence me”
Use	“used the others”	“sometimes used the others”	“often used the others”	“did not use the others”
Care and majority	“care about majority”	“sometimes cared about majority”	“often cared about majority”	“did not care about the majority”
Care and others	“care about others”	“sometimes cared about others”	“often cared about others”	“did not care about the others”

Note. The keywords used for labeling the strategies description given as free-text by participants.

(*Appendix continues*)

Table A4
Keywords Used in Solo Condition

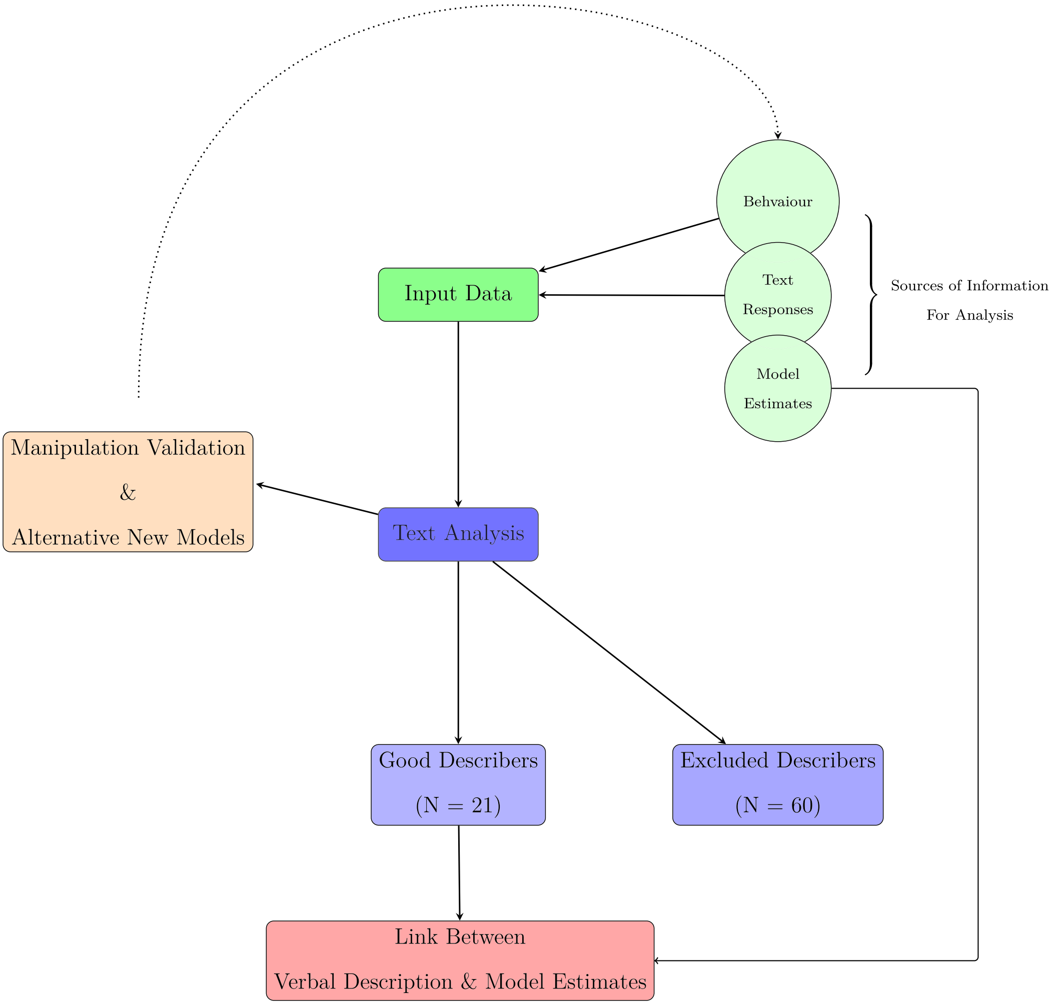
Class	Subclass	Label I	 II	 III	 IV	 V
Jar B	Chose Jar B	"chose Jar B"	"stick to Jar B"			
	Balls	"more black balls"	"jar with the most black"	"higher chance to get black ball"		
	Chance to win	"high *p*(*win*)"	"higher odds of success"			
	Reward	"get 6"				
Jar A	Chose Jar A	"chose Jar A"	"stick to Jar A"			
	Balls	"more white balls"	"higher chance to get white ball"	"jar with the most white"		
	Chance to win	"high *p*(*reward*)"	"riskier higher reward"	"higher value"	"higher chance to get reward"	"greater bonus"
	Reward	"go for £14 bonus"	"get 14"			
Random		"gut feeling"	"intuition"	"just stick"	"just switch"	
Unclassified		"how many black and white balls"	"correct"			

Note. The keywords used for labeling the strategies description given as free-text by participants in the solo condition.

(*Appendix continues*)

Figure A5

Our Conceptual Framework for the Link Between Verbal Description Analysis and Cognitive Models Applied to the Present Case Study


Behvaiour
Text Responses
Model Estimates
Sources of Information For Analysis
Input Data
Manipulation Validation & Alternative New Models
Text Analysis
Good Describers (N = 21)
Excluded Describers (N = 60)
Link Between Verbal Description & Model Estimates

Note. The model starts with input data consisting of three (correlated) sources of information (green circles). To identify good verbal descriptions, the verbal reports are evaluated via text analysis models and are matched with behavior (blue rectangles). *Good describers* are participants, whose verbal descriptions could be successfully classified based on the NLP model output. Once text analysis is performed and good describers are selected (here a total of 21 out of 81 participants), both validation on participants' understanding of the experimental manipulation and new ideas for potential models could be generated (orange rectangle). Finally, the link between the individual NLP model outputs and the individual model fits are correlated (red rectangle). NLP = natural language processing. See the online article for the color version of this figure.

(*Appendix continues*)

Table A5
Individual Raw Verbal Descriptions, NLP Output and Exclusion Criteria

Pid	Verbal description	Risk pref. solo	Best label	Label score	Stage 1—Included?	Stage 2—Included?
1	By how many black or white balls it looked like there were	0.38	How many black and white balls	.97	No	No
2	Looked at how many blacks there were in grid and aimed for the jar with the most black.	0.08	Jar with the most black	.92	**Yes**	**Yes**
3	Estimated the number of black balls.	0.25	Correct	.44	No	No
4	Based on the average number of black balls. From 0 to 49 Jar A and from 50 to 100 Jar B.	0.0	How many black and white balls	.23	No	No
5	What looked correct	0.58	Correct	.96	No	No
6	I played it safe to begin with choosing the outcome with higher odds of success but then if I went on an unsuccessful streak I went for the riskier higher reward for a while.	0.68	Higher chance to get reward	.98	**Yes**	**Yes**
7	Confident in my choices	0.07	Correct	.84	No	No
8	Likelihood of the white balls	0.31	Higher chance to get white ball	.99	**Yes**	No
9	Looked at the grid	0.41	Correct	.33	No	No
10	I tried to stick if I thought it was Jar b	0.0	Stick to jar B	.97	**Yes**	**Yes**
11	Tended to choose Jar b for the higher percentage chance of pulling a black ball.	0.45	Higher chance to get black ball	.99	**Yes**	**Yes**
12	I usually decided to stick with my initial choice of whichever jar I thought the ball came from. There were pros and cons either way. For example, jar a paid a bigger potential bonus, but with a much lower probability.	0.6	Greater bonus	.66	**Yes**	**Yes**
13	I chose the one that would give me the better chance of a reward	0.43	Higher chance to get reward	.99	**Yes**	No
14	Generally tried to go for £14 bonus	0.73	Higher chance to get reward	.95	**Yes**	**Yes**
15	I tried to go with the higher value	0.65	Higher value	.99	**Yes**	**Yes**
16	I guessed which jar it came from by looking at the number of colored balls	0.0	Intuition	.93	**Yes**	No
17	Honestly they were random or gut feeling	0.33	Gut feeling	.98	**Yes**	No
18	I chose based on how many black and white balls in the sample. Then I mostly stuck because I thought that was the better option.	0.55	How many black and white balls	.97	No	No
19	Tried to see which jar it might have come from and then the odds of it being £14 or £6	0.82	Riskier higher reward	.59	**Yes**	**Yes**
20	Based on how many white balls	0.34	Higher chance to get white ball	.78	**Yes**	No
21	By looking at the split of balls on the board and thinking what one would be most likely to come out next to secure the biggest bonus	0.48	Higher chance to get reward	.93	**Yes**	No
22	Assessment of what's come out of the jar and which one gives the best chance of black ball in next draw	0.0	Higher chance to get black ball	.96	**Yes**	**Yes**
23	Determined which urn by higher % of color, for example, if higher % white then thought it was urn 1. If I thought it was urn 1 but the % laid out were close (e.g., say a 60–40 split) I would switch. If I thought it was urn 1 but not many black had been drawn I would stick. I always tried to draw a ball from urn 1.	0.82	Higher chance to get white ball	.88	**Yes**	**Yes**
24	Intuition	0.53	Gut feeling	.98	**Yes**	**Yes**
25	I chose Jar B if it looked like more balls were black. I eventually decided to start sticking even on Jar A because the potential prize was bigger.	0.08	Higher chance to get reward	.99	**Yes**	No

(*table continues*)

(*Appendix continues*)

Table A5 (*continued*)

Pid	Verbal description	Risk pref. solo	Best label	Label score	Stage 1—Included?	Stage 2—Included?
26	I thought it was probably better to aim to choose from Jar B. Although the reward was less, it was 70%.	0.0	Stick to jar B	.99	**Yes**	**Yes**
27	I chose with the potential to get the £14 ball in all cases. I chose it based on the chance of this in each case.	1.0	Go for £14 bonus	.96	**Yes**	**Yes**
28	I always wanted 14	0.57	Higher value	.97	Yes	**Yes**
29	I weighed up the % of each color ball	0.44	How many black and white balls	.72	No	No
30	Gut feelings, did not overthink anything.	0.48	Gut feeling	.99	**Yes**	**Yes**
31	How many white/black balls had been pulled in the bottom row	0.73	How many black and white balls	.99	No	No
32	Generally aimed for a greater chance of obtaining a reward that is, the £6 (seven in 10) over the riskier £14 (three in 10). The jar choice was an assumption on the majority color I saw before me.	0.16	Higher chance to get reward	.99	**Yes**	No
33	Tried to choose the jar by the number of white or black ball percentage picked out and the stuck to the choice.	0.13	Stick	.97	**Yes**	No
34	By seeing how many black balls were out and directly before it and trying to work out the chance of another.	0.53	Higher chance to get black ball	.46	**Yes**	No
35	Based on the amount of balls in each urn	0.3	Higher chance to get reward	.67	**Yes**	No
36	Tried to estimate likelihood that the sample was from each jar	0.37	Get 6	.32	**Yes**	**Yes**
37	I compared how many black ball came out.	0.44	Jar with the most black	.45	**Yes**	**Yes**
38	Looking at the picks	0.42	Higher chance to get reward	.54	**Yes**	No
39	I went by how many black balls I believe I saw and worked on the greater chance of the black ball coming from pot b	0.42	Higher chance to get black ball	.98	**Yes**	**Yes**
40	I chose whichever one had the best odds.	0.25	Higher odds of success	.94	**Yes**	**Yes**

Note. Bolded "yes" entries indicate inclusion of participants in the corresponding stage of the exclusion process. Participants marked with two bolded "yes" entries were included in the final analysis. Verbal descriptions provided by participants on their strategies used to make decisions in the *solo condition* in Experiment 2. *Pid* represents the participants ID. *Question solo condition* are verbal descriptions participants provided about how decisions were made in the *solo condition*. *Risk pref.* represents the proportion of risky choices across all trials in the *solo condition*. *Best label* represent the label that obtained the highest accuracy score by the BERT model. *Label score* represent the score associated with the best label obtained by the BERT model. *Stage 1—Included?* represents whether the individual was included after the first stage of exclusion (in which none classifiable labels were identified). *Stage 2—Included?* represents whether the individual was included after the second stage of exclusion, in which labels that were unsuccessfully matched with choice behavior were excluded. pref. = preference; NLP = natural language processing; BERT = Bidirectional Encoder Representations from Transformers.

Raw Data

In the sections that follow, we present the individual raw verbal descriptions from Experiments 2 and 3, the outputs generated by our natural language processing (NLP) model, and the criteria employed for exclusion.

(*Appendix continues*)

Table A6
Individual Raw Verbal Descriptions, NLP Output, and Exclusion Criteria

Pid	Verbal description	Risk pref. solo	Best label	Label score	Stage 1—Included?	Stage 2—Included?
1	I made decisions based on the likelihood of getting the £6.	0.2	Get 6	0.98	Yes	Yes
2	I thought about probabilities THROUGHOUT	0.09	Higher odds of success	0.44	Yes	Yes
3	I tried to make sure the final pick came from Jar B which would give me the greater chance of a nonzero number.	0.0	Stick to Jar B	0.99	Yes	Yes
4	I tried to choose from Jar B the whole time.	0.4	Chose Jar B	0.89	Yes	Yes
5	Mostly chose to stick	0.8	Chose to stick	0.97	Yes	No
6	I tried to choose the 14£ reward	0.79	Go for £14 bonus	0.98	Yes	Yes
7	Based on what the jar with the highest number of black balls.	0.2	Jar with the most black	0.98	Yes	Yes
8	I guessed the ones with more black were from Jar B	0.28	Odds of black	0.96	Yes	Yes
9	I just tried to figure the odds if it was black I would try to stick and if it was white then I figure most likely will be 0 anyway so I may switch or keep it the same.	0.45	Odds of black	0.96	Yes	Yes
10	I went for the risk and high reward Jar A	0.84	Higher odds of success	0.99	Yes	No
11	I looked at the probability of choosing a black ball for each jar.	0.23	Likelihood to get black ball	0.98	Yes	Yes
12	I chose to stick if I was reasonably sure it was Jar B since Jar B had a higher chance of getting a black ball.	0.0	Higher chance to get black ball	0.99	Yes	Yes
13	By instinct	0.42	Gut feeling	0.99	Yes	No
14	I tried to ensure I was picking Jar B each time	0.0	Stick to Jar B	0.99	Yes	Yes
15	I almost always switched to B, because it had the higher percentage of black balls.	0.06	Higher chance to get black ball	0.99	Yes	Yes
16	By how likely I felt I would get a black ball	0.46	Likelihood to get black ball	0.98	Yes	Yes
17	Based on the sample	0.44	Chose to stick	0.86	Yes	No
18	Whether more balls were black or white	0.06	How many black and white balls	0.94	No	No
19	By looking at the likelihood of a black ball being pulled out	0.05	Likelihood to get black ball	0.98	Yes	Yes
20	When I chose alone, I looked at the balls in the grid and tried to see how many white versus black balls there were; if I felt that there were a lot of white balls, I chose A and then when I saw that there was a lot of black balls, I chose B. I usually stayed with the jar that I felt was whichever one had the most balls, I do not really think I switched much.	0.05	Chose to switch	0.7	Yes	No
21	First, I tried to choose the jar with the most expensive black balls. When I realized that was not working and that I should go with the lower cost black balls of which there were many I did that.	0.4	Chose to switch	0.99	Yes	No
22	Going for the more likely £6.	0.15	Higher value	0.99	Yes	No
23	By looking for what color the majority of the balls were and then when it came to sticking or switching I would make that choice to pick B majority of the times as it would yield a higher probability of a reward	0.3	Higher chance to get reward	0.99	Yes	No
24	I wanted Jar B	0.0	Chose jar B	0.99	Yes	Yes

(*table continues*)

(*Appendix continues*)

Table A6 (*continued*)

Pid	Verbal description	Risk pref. solo	Best label	Label score	Stage 1—Included?	Stage 2—Included?
25	I chose stick or switch depending on which answer I thought would let me pick from Jar B. I tried to choose from Jar B as much as possible since it was more likely to reward me.	0.25	Higher chance to get reward	0.99	Yes	No
260	The choices were made based on probability that there was going to be a black ball, almost exclusively choosing out of Jar B	0.18	Chose Jar B	0.99	Yes	Yes
27	Based upon what I thought the probability would be that a black ball would be picked and, if possible, from Jar A	0.84	Likelihood to get black ball	0.98	Yes	No
28	I always went for Jar B as it had the higher chance of picking a winning black ball	0.0	Higher chance to get black ball	0.99	Yes	Yes
29	Probability of getting a prize was higher in box B so more likely to choose that	0.35	Higher chance to get reward	0.99	Yes	No
30	I wanted the biggest payout, regardless of the chance of getting 0—if I have the option to get the higher payout, I risked it.	1.0	Riskier higher reward	0.99	Yes	Yes
31	I tried to choose B because even though the expected values of Jars A and B were the same (4.20 pounds), I would rather have the greater chance of getting the lesser versus the greater chance of getting nothing.	0.0	Chose jar B	0.98	Yes	No
32	I figured out what would be the most logical outcome and what was best for me.	0.46	Correct	0.96	No	No
33	I prefer the larger chance at a smaller reward so I tried to make sure the draw came from B	0.0	Higher chance to get reward	0.98	Yes	No
34	Chance	0.64	Higher odds of success	0.95	Yes	No
35	By what I felt was the correct answer	0.45	Correct	0.96	No	No
36	I calculated the probability of the Jar using the 10 × 10 grid and then made a decision to go for Jar B that had a higher probability of hitting a black ball albeit at a lower payoff.	0.08	Higher chance to get black ball	0.99	Yes	Yes
37	I made choices based on what balls had been drawn. If it looked like there were a lot of white balls in the draw, I would switch. If the last few balls drawn were white, I would vacillate between switching or sticking.	0.33	Likelihood to get white ball	0.97	Yes	No
38	Random	0.6	Gut feeling	0.91	Yes	No
39	I tried to go for the safest option. If I believed the ball was from jar B, I would always stick with it.	0.09	Stick to jar B	0.98	Yes	Yes
40	With math.	0.76	Correct	0.66	No	No
41	Based off intuition/counter intuition	0.2	Gut feeling	0.99	Yes	No

Note. Verbal descriptions provided by participants on their strategies used to make decisions in the *solo condition* in Experiment 3. *Pid* represents the participants ID. *Question solo condition* are verbal descriptions participants provided about how decisions were made in the *solo condition*. *Risk pref.* represents the proportion of risky choices across all trials in the *solo condition*. *Best label* represent the label that obtained the highest accuracy score by the BERT model. *Label score* represent the score associated with the best label obtained by the BERT model. *Stage 1—Included?* represents whether the individual was included after the first stage of exclusion (in which none classifiable labels were identified). *Stage 2—Included?* represents whether the individual was included after the second stage of exclusion, in which labels that were unsuccessfully matched with choice behavior were excluded. pref. = preference; NLP = natural language processing; BERT = Bidirectional Encoder Representations from Transformers.

(*Appendix continues*)

Table A7
Experiment 1: Individual Raw Verbal Descriptions, NLP Output, and Exclusion Criteria

Pid	Verbal description	Best label	Score	Cog. model fit	Behavior distance
1	By looking how many black and white balls there were and sometimes taking advice from others	Sometimes influenced me	.99	0.0	.12
2	I did not find this information useful but did still look at it	Did not influence me	.99	0.76	.46
3	Still used my own judgment mostly but checked out other peoples and went with them a few times.	Sometimes influenced me	.99	0.0	.25
4	I paid no attention to other people's information	Did not influence me	.99	0.04	.05
5	I went with what looked like the majority went with	Influenced by	.99	0.48	.005
6	I checked the other people's answers about half of the time and the other half I went with trying to be cautious but then going for the riskier option if the cautious answers were not paying off.	Sometimes influenced me	.99	0.0	.04
7	I found them confusing. sometimes I changed my mind based on them and felt like I lost because of it.	Sometimes influenced me	.99	0.61	.37
8	I still went with what I thought rather than the majority	Did not care about the majority	.99	0.0	.21
9	Usually followed the majority	Often cared about majority	.99	0.76	.116
10	I kept my choices the same as the first experiment. Trying to stick when I thought it was jar b because it had more chance of money, even though it was a lesser amount	Sometimes influenced me	.94	0.0	.0
11	Sometimes followed the crowd opinion but mostly favored Jar B still.	Sometimes influenced me	.99	0.0	.11
12	I typically took other people's information into considered and often switched my choice to correspond with the majority. During this round, I often preferred to pick Jar B as that had a higher probability of paying out a bonus overall.	Sometimes influenced me	.99	1.0	.29
13	I tried to find patterns	Sometimes influenced me	.98	0.78	.008
14	Mostly went for the £14 bonus, but did occasionally try doing the opposite to the majority as well as the same	Sometimes influenced me	.99	0.16	.15
15	I tried to go with the higher value	Influenced by	.99	0.0	.34
16	I guessed which jar it came from based on the color of the balls, the other people's information did not really influence me	Did not influence me	.99	0.0	.0
17	Random and gut feeling	Influenced by	.99	0.37	.16
18	The same, I did not find the other people's information persuasive, they have no more data that I do.	Did not influence me	.99	0.07	.11
19	Tried to guess what jar it came from and then went with the majority decision to stick or twist most times	Sometimes influenced me	.98	1.0	.29
20	Looked at how many people stuck or switched to see if my pick was wrong	Influenced by	.99	0.34	.11
21	I considered the majority choice of others but sometime went with my own intuition	Sometimes influenced me	.99	0.28	.1
22	Used the same basic process as when choosing alone but did a cross check to see what other people were doing for additional support	Sometimes used the others	.96	1.0	.50
23	Exactly the same strategy, but I looked at the other choices and it seemed as if most were trying to draw from urn 2??	Influenced by	.97	0.0	.11
24	Influenced by them	Influenced by	.99	1.0	.05

(*table continues*)

(*Appendix continues*)

Table A7 (*continued*)

Pid	Verbal description	Best label	Score	Cog. model fit	Behavior distance
25	I was not interested in other people's guesses as they had no more information than I did. I used the same strategy as the solo mode.	Did not care about others	.99	0.9	.45
26	Often I found going with the majority seemed to lead to me getting rewards more often, so I often used their choices to make mine	Sometimes influenced me	.99	0.99	.48
27	I did exactly the same as I did previously. I did not take any notice of the other people's information.	Did not influence me	.99	0.0	.0
28	Same as without advice	Did not influence me	.99	0.6	.32
29	I did the same as in the choosing alone section, then looked for confirmation bias	Influenced by	.98	0.96	.21
30	Sometimes I went with the information but again gut lead the way.	Sometimes influenced me	.99	0.75	.077
31	I used mostly my original choice but did on occasion take the others into consideration	Sometimes influenced me	.99	0.92	.18
32	Did not change my view greatly and still went with my assessment above. I backed my view over the crowd.	Did not influence me	.99	0.14	.045
33	I got confused on the second part and at one point started to make a different choice to what seemed the logical one.	Sometimes influenced me	.99	0.83	.38
34	Part way through I switched to mainly clicking the same one as the least people picked as this seemed right most of the time unless I really was adamant my choice was right.	Sometimes influenced me	.99	0.01	.086
35	Based on the amount of stick or switches	Influenced by	.99	0.48	.02
36	Same as when I chose alone	Did not use the others	.99	0.91	.061
37	I mostly went with the overall percentage	Influenced by	.99	1.0	.024
38	Still looking at the picks	Influenced by	.99	1.0	.48
39	Same as when I was on my own.	Influenced by	.97	0.51	.33
40	I generally went with what others chose because I noticed that when I played the odds it seemed off so I kind of lost interest in focusing on that.	Sometimes used the others	.99	1.0	.2

Note. Verbal descriptions provided by participants on their strategies used to make decisions in the *solo condition* in Experiment 2. *Pid* represents the participants ID. *Best label* represents the label that obtained the highest accuracy score by the BERT model. *Score* represents the score associated with the best label obtained by the BERT model. *Cog. model fit* represents the individual fit of the *distance* model, and *behavior distance* represents the discrepancy between the NLP model's score and actual behavior. NLP = natural language processing; Cog. = cognitive; BERT = Bidirectional Encoder Representations from Transformers.

(*Appendix continues*)

Table A8

Experiment 3: Individual Raw Verbal Descriptions, NLP Models' Output, and Exclusion Criteria

Pid	Verbal description	Best label	Score	Cog. model fit	Behavior distance
1	I continued to make choices which I thought would secure me the £6, however, there were occasions where I was swayed to go with jar, that is, chance of £14 based on others switching to it.	.99	Sometimes influenced me	0.09	.111
2	U tried not to take notice of others, it made me lose confidence in my own judgment when I paid attention. I eventually decided I was better off ignoring others and I think it went slightly in my favor when I did	.98	Sometimes influenced me	0.32	.118
3	I tried to choose the jar that gave me the largest chance of a nonzero value	.99	Influenced by	0.13	.069
4	It did not affect my choice	.99	Did not influence me	0.0	.333
5	Chose based on the majority but still went mostly with stick	.99	Influenced by	0.48	.233
6	Mostly ignored others information, I had my goal in mind	.99	Did not influence me	0.97	.258
7	I was influenced by what others had done.	.99	Sometimes influenced me	0.61	.127
8	I continued using my previous strategy and ignored what other people choose.	.99	Did not influence me	0.0	.189
9	I tried same strategy as alone, but once seen that was not really working just tried to go with what the majority of 10 people did. If it was tied at 5, I usually just stuck to the stick option.	.99	Sometimes influenced me	0.52	.018
10	I did not let the choices of others sway my opinion, I went with a mixed based on what I fancied	.99	Did not influence me	0.48	.091
11	I looked at the probability of choosing a black ball for each jar. The other people's information did not sway me very much.	.99	Did not influence me	1.0	.319
12	My method of choosing did not change once I had information about other people's decisions.	.99	Did not influence me	0.32	.164
13	I mostly followed my instinct, but on occasion I weighed the answers of others	.99	Sometimes influenced me	0.0	.017
14	Other people's information had little effect and I still tried to choose Jar B each time	.99	Did not influence me	0.33	.176
15	It did not change how I chose, I ignored what the others chose.	.99	Did not influence me	0.2	.053
16	The same as without others choices	.97	Did not influence me	0.31	.068
17	Based on the sample but sometimes, I checked what other people did	.99	Sometimes influenced me	0.79	.358
18	If I thought it was B, I stayed put. If I thought it was A, I switched	.76	Sometimes cares about others	0.13	.044
19	It did not really influence my strategy.	.99	Did not influence me	0.86	.455

Note. Verbal descriptions provided by participants on their strategies used to make decisions in the *solo condition* in Experiment 2. *Pid* represents the participants ID. *Best label* represents the label that obtained the highest accuracy score by the BERT model. *Score* represents the score associated with the best label obtained by the BERT model. *Cog. model fit* represents the individual fit of the *distance* model, and *behavior distance* represents the discrepancy between the NLP model's score and actual behavior. NLP = natural language processing; Cog. = cognitive; BERT = Bidirectional Encoder Representations from Transformers.

(*Appendix continues*)

Table A9

Experiment 3: Individual Raw Verbal Descriptions, NLP Models' Output, and Exclusion Criteria—Continued

Pid	Verbal description	Score	Best label	Cog. model fit	Behavior distance
20	"When I was making choices with the other people's information, I would switch sometimes, when I felt that there was maybe it was a different jar. Sometimes, I saw that only three people would switch, so I ended up taking a chance and switching and actually ended up getting the black ball in Jar A at least 5 times. which is a lot more than I thought I would get and this happened every time I did this strategy. Usually stayed with the one that I picked if I felt like more people switched than stuck with the jar. But it was only when I felt like. Maybe there was a chance to risk, or when I felt like I did not want to risk it and stick with my jar."	.99	Sometimes influenced me	0.68	.306
21	Other people's choices did not matter to me. I trusted my own decision making skills.	.99	Did not influence me	0.0	.378
22	Still going for the more likely £6.	.99	Influenced by	0.16	.059
23	The same as when I was alone but on the odd occasion I was uncertain I would look at the other people's information	.99	Sometimes influenced me	0.33	.191
24	I wanted Jar B	.99	Influenced by	0.57	.305
25	"Other people's information did not affect my choices, I still tried to choose from Jar B as much as possible"	.99	Did not influence me	0.0	.236
26	"Mostly based out of probability, but sometimes when there was a significant amount of people choosing from Jar A, I thought it might be a good idea and it worked out a few times."	.99	Sometimes influenced me	0.28	.055
27	"As above, based upon what I thought might be the probability for a black ball from Jar A (best choice) would be picked—if not high probability then from Jar B. Was helped a little by others' choices at first but saw they were pretty useless: ("	.98	Sometimes influenced me	1.0	.47
28	I ignored the additional information and stuck with my original strategy	.99	Did not influence me	0.14	.067
29	Sometimes looked to see how others decided and sometimes went with majority	.99	Sometimes influenced me	0.15	.157
30	The same as before. The responses were obviously autogenerated and NOT real people because they did not make any sort of sense. I ignored them completely.	.99	Did not influence me	0.034	.018
31	"I tried to ignore the other people, but it sure seemed like the experimenters were toying with me in this section. It seemed like I got a lot more 0's than I would have by the rules as presented. "	.990	Sometimes influenced me	0.078	.036
32	I took into account what people did and made my decision partly based on that.	.990	Sometimes influenced me	0.27	.079
33	Exactly the same way. I tried to get a draw from B.	.98	Influenced by	0.0	.0
34	Follow the crowd (mostly)	.99	Often used the others	0.26	.202
35	By sticking to my guns	.92	Did not influence me	0.14	.176
36	"Well, most of the additional data was 50–50 so that did not help much. If it was a borderline A or B, I went with what the majority wanted to do."	.99	Did not influence me	0.25	.215

(*table continues*)

(*Appendix continues*)

Table A9 (*continued*)

Pid	Verbal description	Score	Best label	Cog. model fit	Behavior distance
37	I used the exact same tactics. What other people chose did not concern me at all.	.99	Did not influence me	0.075	.127
38	I was influenced by others choices	.99	Sometimes influenced me	1.0	.055
39	I did not pay too much attention to their choices to be honest. I was always stuck with my strategy.	.99	Did not influence me	0.0	.018
40	With math	.98	Influenced by	0.47	.07
41	Same way as alone	.99	Did not use the others	0.70	.254

Note. Verbal descriptions provided by participants on their strategies used to make decisions in the *solo condition* in Experiment 3. *Pid* represents the participants ID. *Best label* represents the label that obtained the highest accuracy score by the BERT model. *Score* represents the score associated with the best label obtained by the BERT model. *Cog. model fit* represents the individual fit of the *distance* model, and *behavior distance* represents the discrepancy between the NLP model's score and actual behavior. NLP = natural language processing; Cog. = cognitive; BERT = Bidirectional Encoder Representations from Transformers.

Received May 30, 2023
Revision received April 10, 2024
Accepted April 30, 2024 ■