Verbal Reports Reveal Strategies in Multiple-Cue Probabilistic Inference

Matthew M. Walsh (matthew.walsh.15.ctr@us.af.mil)
Michael Collins (michael.collins.74.ctr@us.af.mil)
Kevin A. Gluck (kevin.gluck@us.af.mil)
Air Force Research Laboratory, 2620 Q Street, Building 852
Wright-Patterson Air Force Base, OH 45433

Abstract

In multiple-cue probabilistic inference, people choose between alternatives based on several cues, each of which is differentially associated with an alternative's overall value. Various strategies have been proposed for probabilistic inference. These include *heuristics*, simple strategies that ignore part of the available information to make decisions more quickly and with less effort. Heuristic models seek to explain the sequence of cognitive events that occur as people make decisions. Validating these models involves evaluating their predictions concerning both outcomes and process measures. In this study, we gathered verbal protocols from participants as they performed multiple-cue probabilistic inference. We find converging evidence across decisions, search behavior, and verbal reports that many participants use a simplifying heuristic, take-the-best. These results provide novel evidence for take-the-best as a process model of human decision behavior in multiple-cue probabilistic inference.

Keywords: Multiple-cue probabilistic inference, verbal protocols, take-the-best, tally, weighted additive

Introduction

Traditional utility theories postulate a decision maker with unlimited time, unlimited mental resources, and complete information about the choice problem (von Neumann & Morgenstern, 1944). This idealization is overly optimistic. In reality, choices must be made quickly, with finite mental resources, and with incomplete information. One way to cope with these challenges is to use simple strategies, or *heuristics*, rather than complex analyses to decide (Simon, 1955). Heuristics are fast because they use simple mental operations, they are frugal because they require little information to enact, and they are surprisingly accurate.

Gigerenzer and colleagues have proposed that the mind contains an adaptive toolbox (2011). The toolbox is comprised of heuristics, their basic building blocks (e.g., search rules, stopping rules, and decision rules), and the cognitive capacities they exploit (e.g., associative memory). This view, though influential, is controversial (see Todd & Gigerenzer, 2000, open peer commentary). The questions of whether heuristics are fast, frugal, and accurate, are separate from the question of whether they adequately describe the sequence of psychological events that occur as people make decisions. In this paper, we examine the last question. We focus on one type of problem, multiple-cue probabilistic inference, and on one heuristic, take-the-best (TTB). Given the centrality of TTB to the adaptive toolbox theory it is important to ask, what is the empirical evidence for TTB as a process model of human decision making?

Multiple-Cue Probabilistic Inference

In multiple-cue probabilistic inference, people decide which of two alternatives has greater value based on multiple cues. Each cue is differentially associated with an alternative's overall value. For example, an investor might consider multiple financial indicators before deciding which of two stocks to purchase. Multiple-cue probabilistic inference is complicated by the fact that no cue or combination of cues perfectly predicts the correct alternative.

Various strategies have been proposed for probabilistic inference.[1] These differ in how many cues they require to enact, and in how they weight each cue. Weighted-additive (WADD) computes the sum of cue values weighted by their importance, and selects the alternative with the greatest resulting value. Tally (TAL) simply selects the alternative with more positive cues. Lastly, take-the-best (TTB) searches cues in order of their validity, and selects an alternative based on the first discriminating cue. TTB is fast because it uses simple mental operations, and it is frugal because it requires little information to enact. Surprisingly, despite its simplicity, TTB often performs as well as – or better than – WADD (Gigerenzer & Gaissmaier, 2011).

Experiment Motivation

Although statistical analyses and computer simulations have demonstrated the feasibility of TTB, they do not establish that people actually use TTB. Doing so requires showing that (1) pre-decisional behavior and (2) choices are consistent with TTB. Empirical results that bear on these issues are mixed. Some studies find that people acquire information in the manner prescribed by TTB, but others do not (Mata, Schooler, & Rieskamp, 2007; Newell & Shanks, 2003). Likewise, some, but not all studies find that the majority of people's decisions are consistent with TTB (Bröder, 2000; Newell & Shanks, 2003). These findings hold even in environments that decidedly favor TTB.

Though informative, the process and outcome measures used to study probabilistic inference have limitations. Search behavior shows what information people acquire and the order in which they do so, but not whether or how they use that information. Additionally, choices predicted by TTB, WADD, and TAL overlap considerably, rendering decisions inconclusive as evidence for process.

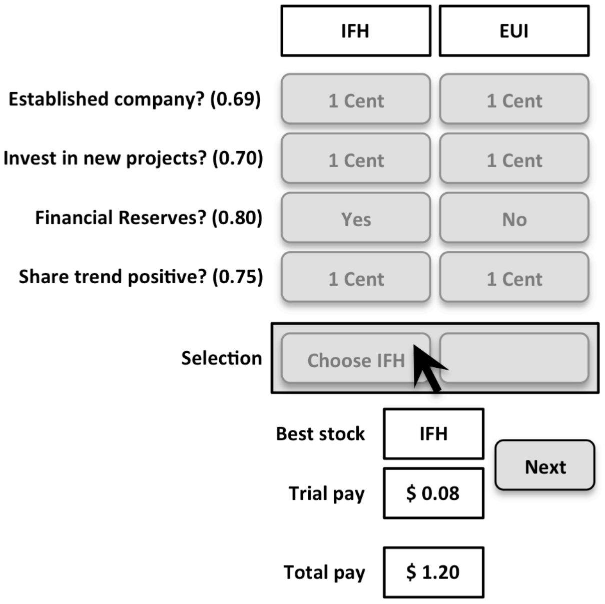
To overcome these limitations, we used verbal protocols

[1] These strategies assume that cue validities are known.

to study the cognitive processes underlying multiple-cue probabilistic inference. Central to protocol analysis is the idea that people can verbalize thoughts, and that mental operations can be inferred from verbalizations (Ericsson & Simon, 1993). Heuristics for probabilistic inference have been formalized as process models (Gigerenzer & Gaissmaier, 2011). These models predict the content of participants' verbalizations and their overt behaviors. Although verbal protocols have not yet been used to study heuristics in multiple-cue probabilistic inference, they have been applied to other topics in judgment and decision making research (Ericsson & Simon, 1993). In the context of this task, verbal data can go beyond search and outcome measures by revealing *how* people use information they acquire to decide. As such, verbal protocol analysis is an appropriate and overdue methodology to incorporate into the scientific study of multiple-cue probabilistic inference.

Experiment

Nineteen people from the University of Dayton participated in a one-hour experiment for monetary compensation. They completed a simulated stock-market selection task that has been used before to study multiple-cue probabilistic inference (Newell, Weston, & Shanks, 2003). In each trial, participants chose between two hypothetical stocks. Each stock had four concealed indicators, which participants could reveal to help guide their decisions (Fig. 1). When the participant clicked an indicator for a stock, the value yes appeared (*Yes*, company IFH has financial reserves), or the value no appeared (*No*, company EUI does not have financial reserves). The value remained visible for the remainder of the trial. To choose a stock, the participant clicked the stock's button below the grid. The name of the best stock and trial pay then appeared.



	IFH	EUI
Established company? (0.69)	1 Cent	1 Cent
Invest in new projects? (0.70)	1 Cent	1 Cent
Financial Reserves? (0.80)	Yes	No
Share trend positive? (0.75)	1 Cent	1 Cent
Selection	Choose IFH	

Best stock IFH

Trial pay $ 0.08

Total pay $ 1.20

Figure 1. Experiment interface. Stock names appeared at the top of the screen, and indicator labels and cue validities were displayed along the left-hand side of the screen. Indicator values were concealed in the gray grid.

Trial pay depended on two factors. First, participants received ten cents for selecting the winning stock and zero cents otherwise. Second, participants paid one cent to view each indicator. To maximize trial pay, they needed to view enough indicators to make informed decisions without spending too much acquiring information. Cue validities were freely visible throughout the experiment, and were identical to those used in other multiple-cue probabilistic inference experiments (Bröder, 2000; Newell, Weston, & Shanks, 2003). Assignment of cue validities to indicators and screen locations was constant throughout the experiment and varied across participants. The payoff structure of the task was explained to participants before the experiment, as was the meaning of cue validity.

The experiment contained 120 trials. Cue configurations were created such that TTB and WADD predicted different choices from one another in 20 trials (16%), and that WADD and TAL predicted different choices from one another in 30 trials (25%).

Participants were told to think aloud during the experiment. Prior to beginning the experiment, they received instruction about how to provide verbal reports, and they practiced verbalizing in three warm-up tasks that were not related to the main task (Ericsson & Simon, 1993; Fox, Ericsson, & Best 2011). Following training, participants received instruction about the experiment. If a participant was silent for longer than one trial, they were reminded to continue thinking aloud.

Results

Behavioral Outcomes

On average, participants selected the correct stock on 73% (± 1 SE) of trials. They revealed 3.35 cues (± 0.24 SE) and earned 3.97 cents (± 0.18 SE) per trial.

To characterize participants' strategies, we compared their decisions to the predictions of three choice rules: TTB, WADD, and TAL. We used a maximum likelihood approach to calculate the probability of each participant's decisions separately for the three strategies. The strategies prescribe deterministic selections for each trial. To allow for stochasticity, we incorporated application errors in the strategy models (Scheibehenne, Rieskamp, & Wagenmakers, 2013). The probability of the observed choice in trial t (O_t) using strategy i (S_i) was

$$P(O_t|S_i, \varepsilon_i) = P(O_t|S_i) \cdot (1 - \varepsilon_i) + 0.5 \cdot \varepsilon_i.$$

Application error (ε_i) is the probability of applying a strategy incorrectly. Under the assumption that application errors result in random selection (i.e., "tremble errors"; Scheibehenne, Rieskamp, & Wagenmakers, 2013), the probability of the observed choice following an application error is 0.5.

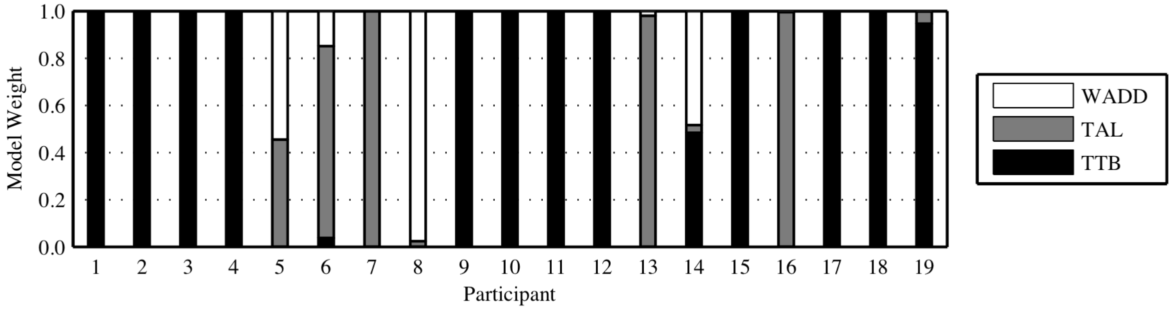


Figure 2. Model weights of three strategies for each participant. Larger values denote greater support for that model.

For each participant and strategy, we estimated the value of application error that maximized the sum of the log-likelihood of the observed choices across the sequence of 120 trials,

$$LLE_i = \sum_{t=1:120} ln\big(P(O_t|S_i, \varepsilon_i)\big).$$

We converted the likelihoods into model weights (Liu & Smith, 2009). Weights add up to one across the three strategies for each participant. Values near zero indicate little evidence for that strategy, and values near one indicate substantial evidence for that strategy.

Fig. 2 shows model weights for each participant. Weights provided greatest evidence for TTB in most participants (12 total). For fewer participants, weights provided greatest evidence for WADD (2 total) or TAL (4 total). On average, participants adhered to the strategy that best fit their choices in 87% of trials.

TTB is a non-compensatory strategy; cues with low validity cannot compensate for the value of a more valid cue. Conversely, WADD and TAL are both compensatory strategies; cues with low validity can compensate for the value of a more valid cue. For this reason, and because model weights only weakly distinguished between them, we combined WADD and TAL into a single category and classified each participant as using a non-compensatory strategy (TTB) or a compensatory strategy (WADD/TAL). Twelve participants used a non-compensatory strategy, six used a compensatory strategy, and one had identical weights for non-compensatory and compensatory strategies.

We compared the performance of participants in the non-compensatory and compensatory groups. Those who used a non-compensatory strategy selected the correct stock slightly more often (74% versus 72% of trials, $t(16) = 0.75$, *n.s.*) and revealed fewer cues (3.1 versus 3.8, $t(16) = 1.29$, $p < .1$). The net effect was that participants who used a non-compensatory strategy earned significantly more per trial (4.2 cents versus 3.3 cents, $t(16) = 2.60$, $p < .05$).

Search Process

Besides specifying decisions, heuristics models predict the order in which information is acquired (*search rules*) and when search stops (*stopping rules*). Specifically, TTB searches cues in order of their validity and stops after finding one discriminating cue. To gather converging evidence for the outcome-based classifications, we examined participants' searching and stopping behavior.

People classified as using a non-compensatory strategy revealed the most-valid indicator first in 92% of trials, while those who used a compensatory strategy only did so in 41% of trials, $t(16) = 6.07$, $p < .0001$. Additionally, people classified as using a non-compensatory strategy stopped searching immediately after finding one discriminating cue in 85% of trials, while those who used a compensatory strategy only did so in 51% of trials, $t(16) = 2.35$, $p < .05$.

Verbal Reports

The experiment produced about 11 hours of verbal data, which were transcribed and segmented into 12,602 task-related utterances. Each utterance was assigned to one of nine categories: (1) search, (2) encoding, (3) single-indicator elaboration, (4) multi-indicator elaboration, (5) unjustified decision, (6) single-indicator decision, (7) multi-indicator decision, (8) feedback evaluation, and (9) metacognitive. One investigator coded 100% of utterances, and a second investigator coded 10% of utterances. The mean inter-rater reliability, measured by Cohen's kappa, was 0.97.

Elaboration and decision statements are especially informative with respect to decision process. Single-indicator elaboration statements compare one indicator between stocks, and single-indicator decision statements base choices on one indicator (Table 1). Conversely, multi-indicator elaboration statements combine information across multiple indicators within a stock, and multi-indicator decision statements base choices on multiple indicators (Table 1). Because TTB involves comparing indicators

Table 1. Examples of Single- and Multi-Indicator Elaboration and Decision Statements

	Single-Indicator	**Multi-Indicator**
Elaboration	*KRL is better on the top indicator*	*I have two yesses and a no for RJB*
Decision	*Take TJM because of share trend*	*Choose this since it's a yes for both*

between stocks and selecting a stock based on the first discriminating indicator, participants using TTB should make more single-indicator elaboration and decision statements. Alternatively, because WADD and TAL involve combining information across multiple indicators within stocks, participants using WADD and TAL should make more multi-indicator elaboration and decision statements.

To test this hypothesis, we calculated the relative proportions of single-indicator elaboration and decision statements – that is the number of single-indicator statements divided by the total number of single- and multi-indicator statements. We compared these proportions between participants who were classified as using a non-compensatory (TTB) or compensatory (WADD/TAL) strategy based on their choices (Fig. 3). Relative to participants who used WADD or TAL, those who used TTB made more single-indicator elaboration ($t(16) = 3.12$, $p < .01$) and decision statements ($t(16) = 4.39$, $p < .001$).

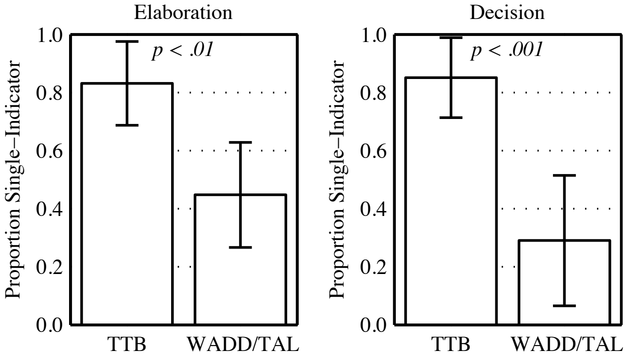


Figure 3. Proportion of single-indicator elaboration and decision statements for participants classified as using a non-compensatory strategy (TTB) or a compensatory strategy (WADD/TAL).

Multi-Dimensional Classification

In the previous sections, we classified participants based on their decisions, and we sought converging evidence in the form of search behavior and verbal reports. This is the predominant approach used in the literature. A potentially more powerful approach is to classify participants based jointly on their decisions, search behavior, and verbal reports – that is, a multi-modal approach (Schulte-Mecklenbeck, Kühberger, & Ranyard, 2011).

To combine information from these three sources, we performed a multi-dimensional classification. For each participant we recorded (1) the model weight assigned to TTB, (2) the probability of terminating search immediately after the first discriminating cue, and (3) the relative proportion of single-indicator statements. To place equal emphasis on the three dimensions, we normalized the values within each using z-scores. We then used a two-step k-means cluster analysis to determine the number of clusters of participants and to assign each participant to a cluster.

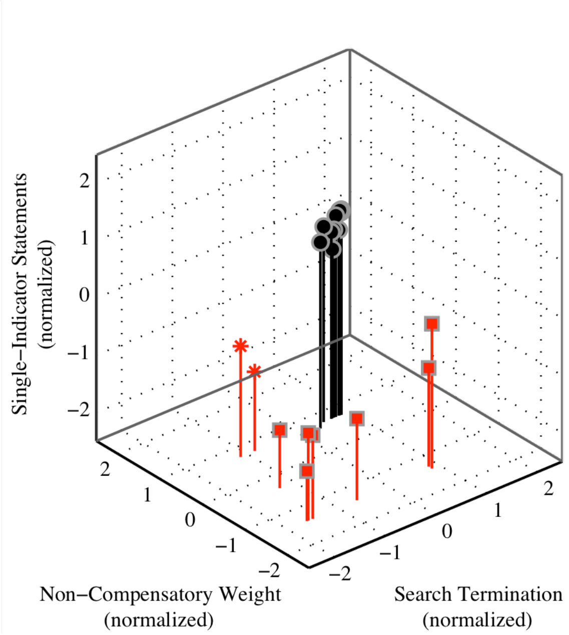


Figure 4. Scatter plot of normalized search measure (termination), outcome measure (non-compensatory decision weight), and verbal measure (proportion of single-indicator elaboration and decision statements). Black circles denote participants who acted according to TTB and red squares denote participants who acted according to WADD or TAL. Red stars denote two participants classified differently by outcome-based and multi-dimensional approaches.

The analysis revealed two clusters with ten individuals and nine individuals (Fig. 4). Participants in black acted most consistently with a non-compensatory strategy: their decisions were consistent with TTB, they stopped searching after finding one discriminating cue, and they predominantly made single-indicator statements.

Participants in red, though somewhat more variable, acted most consistently with a compensatory strategy: their decisions were consistent with WADD or TAL, they continued searching after finding one discriminating cue, and they predominantly made multi-indicator statements.

The results of the multi-dimensional classification matched the outcome-based classification with the exception of the two participants depicted by red stars. These participants' decisions were consistent with TTB, but they continued searching after finding a discriminating cue and they made many multi-indicator statements.

Consistency of Verbal Reports

One concern with verbal protocol analysis is that verbal reports may deviate from behavior (Nisbett & Wilson, 1977; Schooler, 2011). To address whether participants' verbalizations were consistent with their choices, we examined a subset of trials that satisfied two criteria. First, the heuristics needed to predict different decisions based on the set of cue values (16% of trials). Second, the participant needed to make an elaboration or decision statement that was informative with respect to their decision process (33% of trials). The conjunction of these relatively low-frequency events limited our sample to a total of 92 trials across all participants.

This occurred in 52 trials with elaboration statements, and in 41 trials with decision statements. In these trials, verbalizations were overwhelmingly consistent with choices (elaboration statements: 37/52; decision statements: 35/41). Elaboration statements were not as consistent with choices as decision statements (71% versus 85%), $\chi^2 = 2.65$, $p = .05$ (one-tailed).

In a total of 21 trials from 8 different participants, verbalizations and choices were inconsistent. We examined these trials to determine why the verbalizations and choices diverged. We found that most inconsistent trials could be assigned to three categories (Table 2).

Table 2. Categorization of Inconsistent Trials

Category	Number of Trials
Frugal WADD/TAL	8
Separate Search and Decision	6
Guessing	2
Unclassified	5

In eight inconsistent trials, participants appeared to apply WADD/TAL using only a subset of the cues (e.g., two of the cue pairs rather than all four). Because they did not reveal all of the cues, their decisions – which seemed to follow a compensatory process – ended up coinciding with TTB. Had they applied the same decision processes (i.e., counting the number of yesses for each stock) based on all four cues, they would have likely selected the other stock, as predicted by WADD/TAL. We labeled these trials *frugal WADD/TAL*. Seven of the eight trials in this category came from the two participants who were reclassified from TTB to WADD/TAL in the multi-dimensional classification.

In six inconsistent trials, participants examined cues in the order prescribed by TTB, and they made elaboration statements consistent with that strategy. However, they did not terminate search after the first discriminating cue. Rather, they revealed additional cues and decided based on the majority of values for the two stocks. We labeled these trials *separate search and decision.*

In two *inconsistent* trials, participants began by examining cues in order of their validity, but decided before revealing any discriminating cue. That is, they abandoned their strategy mid-trial and *guessed* rather than spending more to find a discriminating cue. In the remaining five *unclassified* trials, the reason for the inconsistency between decisions and reports was unclear.

Conclusions

The two main results of this study can be summarized simply. First, most participants' decisions were consistent with TTB, although a large minority appeared to use WADD or TAL. Second, outcome measures, search behavior, and verbal reports converged. Together, these results strongly support the notion that many individuals adopt TTB in this type of task and environment.

Heuristics models seek to explain the sequence of psychological events that occur as people make decisions. To adequately test these models, one must evaluate their predictions concerning both outcome measures and process measures (Schulte-Mecklenbeck, Kühberger, & Ranyard, 2011). A priori, it was unknown whether verbal protocols, a high-density performance measure, would be consistent with the predictions of TTB. We found that they were in participants who decided in the manner prescribed by TTB. This provides novel evidence for TTB as a *process model* of multiple-cue probabilistic inference.

Verbal Reports as a Source of Process Data

A common concern with verbal protocols is that they may be inaccurate (Nisbett & Wilson, 1977; Schooler, 2011). In fact, verbal reports are less accurate when individuals must retrieve information from long-term memory, and when retrieval is difficult (Ericsson & Simon, 1993). This may be the case with retrospective reports. To avoid this problem, we gathered protocols concurrently with task performance. We observed substantial consistency across outcome measures, search behavior, and concurrent verbalizations. Participants whose decisions were most consistent with TTB also made more single-indicator elaboration and decision statements. The correspondence between verbal reports and decisions held at the level of single trials. When informative verbalizations accompanied diagnostic choices, reports were consistent with 77% of decisions (72/93).

Inconsistent trials, though uncommon, were nonetheless informative. The finding that some participants applied a compensatory strategy to a subset of cues makes the point that variants of WADD and TAL can be used in conjunction with non-exhaustive information search (Schulte-Mecklenbeck, Sohn, de Bellis, Martin & Hertwig, 2013). The majority of frugal WADD/TAL trials (seven out of eight) came from just two participants. These same participants were classified differently based on decisions alone, versus decisions along with search behavior and verbal reports. The reason why these participants were so difficult to classify is because their actual strategies likely fell somewhere between TTB and WADD/TAL, or what we identified above as *frugal* WADD/TAL.

Some participants made non-compensatory elaboration statements but appeared to use a compensatory decision rule. Heuristics are comprised of three basic building blocks: search rules, stopping rules, and decision rules (Gigerenzer & Gaissmaier, 2011). Dissociations between elaboration statements, which occur during the search portion of the trial, and choices reflect the separability of the building blocks; the individual can examine cues in the manner prescribed by one heuristic, but decide according to another. This has led some researchers to question the utility of process measures that focus on search. Indeed, we found that decision statements were more consistent with choices than elaboration statements. That fact notwithstanding, the majority of elaboration statements were consistent with choices as well.

A second concern with protocol analysis is that thinking aloud may interfere with the cognitive processes under investigation (i.e. *reactivity*; Nisbett & Wilson, 1977; Schooler, 2011). To control for this possibility, we gathered data from additional participants in a silent, control condition. There were no differences between groups in terms of information acquisition or decision behavior.

Limitations and Future Directions

According to the adaptive toolbox theory, the mind contains a collection of heuristics, one of which is TTB (Gigerenzer & Gaissmaier, 2011). The adaptive toolbox theory predicts that people will only use TTB when it is adaptive to do so. For example, when cue validities are non-compensatory and when information is costly. These conditions were met in our experiment. Accordingly, more than half of participants used TTB.

The adaptive toolbox theory also predicts that people will use other strategies like WADD and TAL when cues are compensatory and when information is free. Other studies have confirmed these predictions (Bröder, 2003). If we replicated our experiment under such conditions, we expect that more participants would decide according to WADD and TAL, and that the relative frequency of multi-indicator elaboration and decision statements would increase. Although these predictions remain to be tested, participants classified as using compensatory strategies did exhibit the expected verbalization patterns for elaboration and decision statements in this experiment.

Simon (1992) stated that our methods for gathering data must fit the shapes of our theories. The primary innovation of this work is the application of verbal protocols to the study of heuristics in multiple-cue probabilistic inference. Participants' verbalizations provided critical information about how they used the information they acquired to make decisions. Of course, people do not always take-the-best. The method used here, verbal protocol analysis, has considerable potential to enhance understanding of the strategies people use in different environments, for different types of problems, and in different tasks.

Acknowledgments

The views expressed in this paper are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government. This research was performed while M.M.W. held a National Research Council Research Associateship Award with the AFRL's Cognitive Models and Agents Branch. The authors thank AFRL for supporting this research.

References

Bröder, A. (2000). Assessing the empirical validity of the "take-the-best" heuristic as a model of human probabilistic inference. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 1332-1346.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis. Verbal protocols as data*. Cambridge, MA: 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*, 316-344.

Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual Review of Psychology, 62*, 451-482.

Liu, C. C., & Smith, P. L. (2009). Comparing time–accuracy curves: Beyond goodness-of-fit measures. *Psychonomic Bulletin & Review, 16*, 190-203.

Mata, R., Schooler, L. J., & Rieskamp, J. (2007). The aging decision maker: Cognitive aging and the adaptive selection of decision strategies. *Psychology and Aging, 22*, 796-810.

Newell, B. R., & Shanks, D. R. (2003). Take the best or look at the rest? Factors influencing "one-reason" decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 53-65.

Newell, B. R., Weston, N. J., & Shanks, D. R. (2003). Empirical tests of a fast-and-frugal heuristic: Not everyone "takes-the-best". *Organizational Behavior and Human Decision Processes, 91*, 82-96.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231-259.

Scheibehenne, B., Rieskamp, J., & Wagenmakers, E. J. (2013). Testing adaptive toolbox models: A Bayesian hierarchical approach. *Psychological Review, 120*, 39-64.

Schooler, J. W. (2011). Introspecting in the spirit of William James: Comment on Fox, Ericsson, and Best (2011). *Psychological Bulletin, 137*, 345-350.

Schulte-Mecklenbeck, M., Kühberger, A., & Ranyard, R. (2011). The role of process data in the development and testing of process models of judgment and decision making. *Judgment and Decision Making, 6*, 733-739.

Schulte-Mecklenbeck, M., Sohn, M., de Bellis, E., Martin, N., & Hertwig, R. (2013). A lack of appetite for information and computation. Simple heuristics in food choice. *Appetite, 71*, 242-251.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics, 69*, 99-118.

Simon, H. A. (1992). What is an "explanation" of behavior? *Psychological Science, 3,* 150-161.

Todd, P. M., & Gigerenzer, G. (2000). Précis of simple heuristics that make us smart. *Behavioral and Brain Sciences*, *23*, 727-741.

Walsh, M. M., & Gluck, K. A. (2015). Verbalization of decision strategies in multiple-cue probabilistic inference. *Journal of Behavioral Decision Making.*

von Neumann, J., & Morgenstern, O. (1944). *Theory of gambles and economic behavior*. Princeton, NJ: Princeton University Press.