UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

HOW DO SKILLED DECISION MAKERS SOLVE HARD PROBLEMS? A PROTOCOL ANALYSIS OF THE BERLIN NUMERACY TEST

A THESIS

SUBMITTED TO THE GRADUATE FACULTY

In partial fulfillment of the requirements for the

Degree of

MASTER OF SCIENCE

By

LONG V. NGUYEN
Norman, Oklahoma
2025

HOW DO SKILLED DECISION MAKERS SOLVE HARD PROBLEMS?
A PROTOCOL ANALYSIS OF THE BERLIN NUMERACY TEST

A THESIS APPROVED FOR THE
DEPARTMENT OF PSYCHOLOGY

BY THE COMMITTEE CONSISTING OF

Dr. Edward T. Cokely, Chair

Dr. Adam Feltz

Dr. Mauricio Carvallo

Acknowledgement

This work would not have been possible without the help and support of many people. I would like to express my deepest gratitude to Dr. Edward Cokely for his guidance, support, and encouragement throughout this process. I am also grateful to my committee members Dr. Adam Feltz, Dr. Rocio Garcia-Retamero, and Dr. Mauricio Carvallo for their time and constructive feedback. I would like to thank my colleagues at the Decision Analytics Lab for their continued support and motivation. Finally, I am deeply thankful to my friends and family for their unwavering love and support in my academic journey. In particular, I would like to express heartfelt gratitude to my brother and my mother, whose encouragement has inspired me to persevere and pursue higher academic achievements, despite challenges and setbacks.

Table of Contents

List of Tables

List of Figures

Abstract

The Berlin Numeracy Test is one of the strongest predictors of general decision making skill and risk literacy. Despite hundreds of studies linking test performance with decision quality, there is limited evidence on how people solve problems on the test itself. Are skilled decision makers better abstract reasoners who use calculations to solve hard math problems, or are they more like experts who use heuristics to develop a deeper conceptual understanding that can simplify problem solving? To answer this question, I conducted a Protocol analysis on verbalizations expressed during problem solving on the Berlin Numeracy Test (*n* = 160). Results revealed that most people only relied on formal calculations for easy problem solving, switching to heuristic-based strategies for difficult problems. Cognitive path modeling indicated that the relationship between skilled decision making and problem solving was fully mediated by heuristic use, which helped people understand problems in a way that eliminated the need for complex calculations. Results suggest that the predictive power of the Berlin Numeracy Test may be partly explained by this special property of its difficult questions (e.g., computational simplification via conceptual understanding). Results are interpreted in accordance with Skilled Decision Theory (Cokely et al., 2018), emphasizing the role of representative understanding in superior decision making and effective problem solving. Limitations and implications for measurement and training are briefly discussed.

Keywords: verbal protocol analysis, problem solving, heuristics, numeracy, risk literacy

HOW DO SKILLED DECISION MAKERS SOLVE HARD PROBLEMS?
A PROTOCOL ANALYSIS OF THE BERLIN NUMERACY TEST

CHAPTER 1: INTRODUCTION

Skilled decision makers are less biased and more informed about choices in many domains (Cokely et al., 2018; in press; see also Brune de Bruin et al., 2007; Garcia-Retamero et al., 2017; Skagerlund et al., 2018; Stanovich, 2016; Stanovich & West, 2000). Skilled decision makers are also better at mathematical problem solving as measured by numeracy tests (Cokely et al., 2012; Peters et al., 2008; Peters, 2020; Reyna, 2008; Reyna & Brainerd, 2023). For example, they score higher on the Berlin Numeracy Test, which asks participants to solve probabilistic problems that have objectively correct answers such as:

> Imagine we are throwing a loaded die (6 sides). The probability that the die shows a 6 is twice as high as the probability of each of the other numbers. On average, out of these 70 throws how many times would the die show the number 6?

Although there are hundreds of studies on the relations between decision making and numeracy, there is very limited evidence investigating underlying problem solving processes. How do skilled decision makers solve hard numeracy problems?

One way people could solve difficult numeracy problems would be to derive mathematical equations and use formal calculation procedures. This is consistent with leading theories suggesting that effective decision makers may often override their heuristics and biases in favor of more abstract, logical reasoning processes (De Neys, 2006; Kahneman, 2011; Kuhn, 2000; Schirrmeister et al., 2020; Stanovich & West, 1998; 2000; 2008; Toplak et al., 2014). This also accords with evidence indicating that effective problem solvers tend to derive abstract equations and use formal calculations on easy and well-practiced problems (Chi et al., 1981; Hegarty et al., 1995; Mayer et al., 1996; Mayer & Hagerty, 2012). However, it is inconsistent with a large body of evidence indicating that experts usually rely on heuristics and deeper conceptual understanding for difficult problem solving, including difficult mathematical problem solving (Newell & Simon, 1972; see also Chase & Simon, 1973; Polya, 1954; Schoenfeld, 1989).

One source of evidence on expert problem solving is Protocol Analysis, which is a method for collecting and analyzing sequences of verbalized thoughts (Ericsson & Simon, 1980; 1993). Although people cannot reliably explain their own cognitive processes (Nisbett & Wilson, 1977), research indicates they can accurately verbalize the contents of their short-term memories during reasoning (De Groot, 1946; Fox et al., 2011; Newell & Simon, 1972; Payne, 1976). Accordingly, the method trains participants to verbalize their thoughts during reasoning without summarizing or explaining. Verbalizations are then transcribed and coded for quantitative and qualitative analyses of *operators* (e.g., heuristics, calculations, transformations), *states* (e.g., specific configurations of mental problem representations), and *nodes* (e.g., sequences of transition points within a problem space). For example, in one influential study, Chi et al. (1981) used verbal protocol analysis to examine the problem solving strategies of physics experts (i.e., professors) and novices (i.e., college students). Results revealed that experts used self-explanatory techniques, relying less on formulas and more on conceptual understanding (e.g., state transitions reflecting "deep knowledge structures" rather than abstract equations; Chi et al., 1981; 1982). In contrast, novice problem solvers tended to focus on surface-level problem features and procedural methods (e.g., plug and chug equations), which led to poorer problem solving performance on difficult problems.

Verbal protocol analysis has also been used to investigate processes linking higher numeracy scores with risky choices (Cokely & Kelley, 2009). Theoretically, more numerate people could make better risky choices by using formal calculations (e.g., calculating expected values; Frederick, 2005; Stanovich & West, 2000). However, results of a retrospective protocol analysis suggested that superior decision making was explained by heuristic deliberation and personal narratives (e.g., verbalizing things like ‘even though it probably won’t happen, my mom would be so upset if I lost that much money’). Put another way, numerate people seem to use self-explanatory techniques to develop a personalized conceptual understanding of potential consequences and risks (e.g., how different outcomes might feel and why). Given these and other findings, it seems possible that heuristics may often be used by skilled reasoners to develop a more representative understanding of many complex tasks, potentially including difficult numeracy tasks (e.g., Skilled Decision Theory; Cokely et al., 2018; in press). However, I have not been able to find any published research that has directly assessed the cognitive processes skilled decision makers use when solving difficult numeracy problems. Therefore, I conducted a protocol analysis to identify and analyze strategies that were concurrently verbalized during problem solving on the Berlin Numeracy Test.

CHAPTER 2: METHOD

Participants

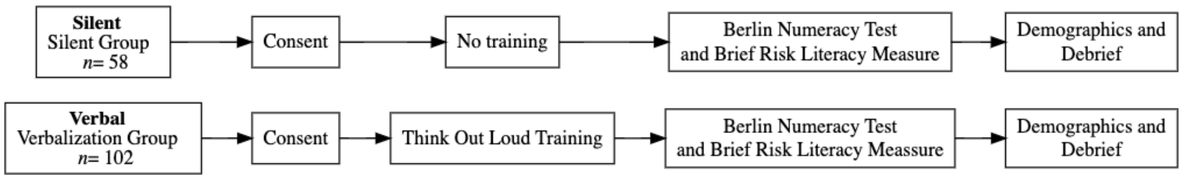
One hundred and seventy-three people initially agreed to participate in the study and were semi-randomly assigned to two groups. When two participants were available at the same time, the group assignment was random. When only one person was available at a specific time they were assigned to the verbal condition to maximize power for the protocol analysis, following common practices (Ericsson & Simon, 1993; Fox et al., 2011). Responses from 13 participants (11 verbal and two silent) were excluded from the final analysis because (a) they did not complete the study, (b) they requested that their data not be used for analysis, or (c) they completed the study unreasonably quickly, defined as more than two standard deviations faster than the group mean (i.e., less than 12 minutes for the silent condition or less than 29 minutes for the verbal condition that included 10-15 minutes of verbal protocol training). In total, data from 160 participants were included in the final analysis (102 verbal and 58 silent). Nearly all participants were young adults ranging from 18-24 years of age,[1] including 41 people who identified as male (28 verbal and 13 silent) and 119 who identified as female (74 verbal and 45 silent).[2]

Procedure

Figure 1

Diagram of the complete procedure during the study

[1] A few participants who were under the age of 18 received parental or guardian permission and university approval to participate in studies at the University of Oklahoma prior to participating in this study.
[2] Due to a programming error, demographic data was collected using two different scales that had slightly different wording (see appendix).



The diagram (Figure 1) provides a summary of the study procedure. Each group completed the study in a separate room. After informed consent was collected, the silent group completed the study alone and was monitored via microphone. The verbalization group had a trained research assistant present throughout the study. The research assistant provided verbalization training and occasionally prompted the participant with reminders to think out loud if they ever became silent or spoke too quietly (Cokely et al., in press; Ericsson & Simon, 1993; see appendix).[3] Both groups subsequently completed the Berlin Numeracy Test and several other cognitive measures, before completing demographics and being debriefed. Demographic questions and results can be found in the appendix.

Measures

Statistical Numeracy

The four-item Berlin Numeracy Test was used to measure the participant's statistical numeracy, which is known to be a robust predictor of decision making quality including people's ability to understand and evaluate risk (i.e., risk literacy; Cokely et al., 2012).

Skilled Decision Making Assessment

To assess skilled decision making, I used the Brief Risk Literacy Test, which includes seven judgment and decision making questions that have been found to have good psychometric

[3] Most training items for Verbal Protocol Analysis are from Cokely and Kelley (2009), which was developed in consultation with K. Anders Ericsson. Minor modifications were made in consultation with Edward T. Cokely.

properties, spanning financial, medical and health topics (e.g., evaluating loan options; selecting high-quality medications; Cokely et al., in press; but for detailed analyses see Allan et al., 2021). As part of a separate protocol analysis study, participants also completed several other tasks that were not part of the current analysis, including the Berlin Numeracy Test Components assessment (Cokely et al., in press; Ghazal et al., 2014), the Schwartz Numeracy test (Schwartz et al., 1997), two additional Risk Literacy Questions from Cho et al. (2024), a Graph Literacy Test (Garcia-Retamero et al., 2016; Okan et al., 2019), six additional Graph Literacy Questions from Cho et al. (2024), a Subjective Numeracy Scale (Fagerlin et al., 2007), and a Subjective Graph Literacy Scale (Garcia-Retamero et al., 2016).

Coding and transcriptions of verbal protocols

Transcription

The recordings of the participants' verbalizations were saved and stored on a secure cloud-based hard drive. For transcription, I used an AI transcription program from OpenAI (2023; 2022) called Whisper, with Python programming software (Python Foundation Software, 2023). This system has been found to provide near-professional grade transcription for similar scientific projects (Ambrogio et al., 2023; Point & Baruch, 2023).

Coding scheme

An *a priori* coding scheme was developed for this study (i.e., mapping the relevant problem space; Newell & Simon, 1972). Explanations and examples of codes and responses for each of the different strategies were provided to both protocol coders (see Table 1). Minor or irrelevant deviations in problem strategies were not considered when coding the protocols for the final analysis (i.e., coding only considered relevant protocols; see Newell & Simon, 1972; Ericsson & Simon., 1980; 1984; 1993). For example, if a participant started by verbalizing some heuristics but then changed strategies and calculated an answer it was coded as calculation. Likewise, if a participant verbalized some calculations but ultimately used heuristics to transform and solve the problem, it was coded as heuristic. When a participant verbalized other ambiguous information without verbalizing specific or discernable sequences of steps it was coded as ambiguous. Table 2 provides examples of the final coding schemes for all problems on the Berlin Numeracy Test.

Protocols for final analysis

After coding, I analyzed the protocols associated with correct answers on the Berlin Numeracy Test. Protocol analysis was conducted separately for the easy and hard problems. For analysis, I defined hard items on the Berlin Numeracy Test as those answered less than 50% of the time (Cokely et al., 2012; Ghazal et al., 2014; Friederichs et al., 2020).

Table 1

Explanation of coding scheme for verbalizations

Type of Coding	Explanation of coding	Response to question About how many times do you expect a coin to come out heads after 1000 flips?	Explanation of response
Heuristic	This term is defined as the use of cues or rule of thumb (influenced by problem environment) instead of the use of formal operational formulas when making decision (Gigerenzer, 2008).	"A coin has two sides, heads or tails, so it is 500"	Knowledge of a coin to solve the problem instead of a formal equation like expected value
Calculation	A form of optimalization to problem solving where the problem-solving processes are not largely influenced by environmental factors, but by the ability to work with ideal formulas when called upon by the problem (Fuchs et al., 2014).	"The probability of a coin landing heads is 1 out of 2, so it is ½ multiplied by 1000 is 500"	Forming an expected value equation to solve the problem considering possibilities of an occurrence instead of using one's knowledge of a coin
Ambiguous	This method indicates that no discernable method can be detected from the verbalization given by the participant.	"500"	No indication mentioned in verbalization of how that person solved the problem

Note: Coding was determined by the protocols that came up with the solution. Protocols that indicated a change in strategy were not considered for coding.

Table 2

Coding scheme for each problem on the Berlin Numeracy Test

Statistical Concept	Berlin Numeracy Test Question	Heuristic Response	Calculation Response
Expected Value	Imagine we are throwing a five-sided die 50 times. On average, out of these 50 throws how many times would this five-sided die show an odd number (1, 3 or 5)?	"It seems like it is over half. Since half is 25. Then it should be 30."	"Since we have three odd numbers and five sides. It should be 3 over 5. 3 over 5 times 50. 3 times 50 is 150 divided by 5 is 30."
Conditional Probability	Out of 1,000 people in a small town 500 are members of a choir. Out of these 500 members in a choir 100 are men. Out of the 500 inhabitants that are not in a choir 300 are men. What is the probability that a randomly drawn man is a member of the choir?	"Well it seems that we have only 400 men in total. 100 out of them are men. So, it's 1 out of 4."	"So we have 1000 people in the town. 500 people in the town are in the choir and out of those 100 are men. So, that is 100 over 1000 for man and choir in town. 500 people in the town are not in the choir and out of those 300 are in the town. So, that is 300 out of 100 man and not choir in town. So altogether, we have 400 men, so the probability of man in the town is 400 out of 1000. Due to conditional probability, event over given. So, 100 over 1000 is event and 400 over 1000 is given. Therefore, 100 over 1000 divided by 400 over 1000 is 100 over 400 or simplified into ¼ or 25%."
Weighed Probability	Imagine we are throwing a loaded die (6 sides). The probability that the die shows a 6 is twice as high as the probability of each of the other numbers. On average, out of these 70 throws how many times would the die show the number 6?	"If this was a normal die, 60 throws would go to six sides make 10 throws per side." "Since six is twice as much, then six appears 20 times."	"If x represents the numbers of throws for a normal side of die and 2x represents the throws of side six since it is not normal. Therefore, all the throws have to equal the total amount. So, 5x (normal sides) plus 2x (non-normal side must equal 70 (total throws). If you add up all the sides together you get 7x equals to 70, then you divide it by 7 and get x equals to 10.

			Since 2x equals the amount of throws required for the non-normal side, then we take 2 times 10 (replacing x) and get 20.”
Inverse Conditional Probability	In a forest, 20% of the mushrooms are red, 50% are brown, and 30% are white. A red mushroom is poisonous with a probability of 20%. A mushroom that is not red is poisonous with a probability of 5%. What is the probability that a poisonous mushroom in the forest is red?	“Since we are looking for poisonous, there are only two categories red and poisonous and non-red and poisonous. Since 20% of 20 looks to be the same as 80% of 5, then it should be about 50%.”	“Let’s first calculate the probability of getting red and poisonous in the forest, so .2 times .2 equals .04 or 4% of mushrooms are red and poisonous. Next, let’s calculate the probability of brown and poisonous, so .5 times .05, which equals to .025 or 2.5% that a mushroom is brown and poisonous. After that let’s calculate the probability that we have white and poisonous, so .30 times .05 which is .015 or 1.5% that a mushroom is white and poisonous. Find the percentage of poisonous mushrooms in the forest, which is 4% plus 2.5% plus 1.5% all together 8%. Conditional probability event over given, therefore event is red and given is poisonous. So, 4% over 8%, which equals to ½ or 50%.”

Note: Responses are determined by heuristics, calculation, or are ambiguous based on the proximity that the response has to the example. All Berlin Numeracy Test problems have a restricted number of ways they can be solved (i.e, problem space).

CHAPTER 3: RESULTS

Did verbalizing affect problem solving performance?

There were no significant differences in accuracy of problem solving between the verbal group and the silent group on the Berlin Numeracy Test ($t(99.90) = 1.03$, $p = 0.30$, $d = .17$).[4] This was also the case when separately considering the easy problem ($t(158) = 0.21$, $p = 0.83$, $d = .04$) or the hard problems ($t(158) = 1.36$, $p = 0.17$, $d = .22$). I also found no significant differences in the quality of decision-making related to verbalization ($t(158) = 0.48$, $p = .63$, $d = .08$). Findings suggest that verbalization did not significantly affect problem solving or decision making performance (Table 3), consistent with previous findings (Fox et al., 2011).

Table 3

Comparison of verbal group versus silent group

Group	Berlin Numeracy Test Score (out of 4)		Easy Problem (out of 1)		Hard Problem (out of 3)		Decision-Making Skill (out of 4)	
	M	*SE*	*M*	*SE*	*M*	*SE*	*M*	*SE*
Verbal	1.02	0.09	0.57	0.05	0.45	.06	4.09	.12
Silent	1.19	0.14	0.59	0.07	0.60	.10	4.19	.18

Note: No significant differences were found between verbal or silent group in any of the four categories

Were verbalized strategies clearly identifiable?

A total of 96% of accurate responses on the Berlin Numeracy Test were coded as either heuristic use or calculation use, with only four percent coded as ambiguous (see Appendix).[5] To evaluate the inter-rater reliability, two trained coders independently coded the same responses

[4] A Welch's t-test was done instead of a student t-test due to issues of homoscedasticity.
[5] Ambiguous protocols were not significantly related to skilled decision making or results on the Berlin Numeracy Test (see Appendix).

from 20 randomly chosen participants. Overall, results suggested that coding was reliable (ϕ = 0.75, $p < .05$) and generally unambiguous.[6]

Did people use calculations to solve the easy problem?

A t-test revealed that problem accuracy on the easy problem was associated with calculation use (M = 0.54, SE = 0.05) rather than heuristic use (M = 0.06, SE = 0.02), ($t(101) = 7.97$, $p < .05$, $d = .79$). Consistent with expectations, results suggested that the easiest Berlin Numeracy Test question was likely to be solved by using formal calculations.

Did people use heuristics to solve hard problems?

A t-test indicated that difficult problem solving was associated with heuristic use (M = 0.38, SE = 0.06), rather than calculation use (M = 0.05, SE = 0.02), ($t(101) = 5.30$, $p < .05$, d = .53). Consistent with expectations, results suggested that the difficult problems on the Berlin Numeracy Test were more likely to be solved by heuristics.

Table 4

Proportion of verbalizations depicting problem solving accuracy by problem solving type

Problem Solving Strategy Type	Berlin Numeracy Test overall		Easy Problem (Berlin Numeracy Test Problem 1)		Hard Problems (Berlin Numeracy Test Problems 2, 3 and, 4)	
	M	*SE*	*M*	*SE*	*M*	*SE*
Heuristics	0.44	0.06	0.06	0.02	0.38	0.06
Calculations	0.59	0.06	0.54	0.05	0.05	0.02

Note: Ambiguous protocols were excluded because they were unrelated to overall problem solving accuracy

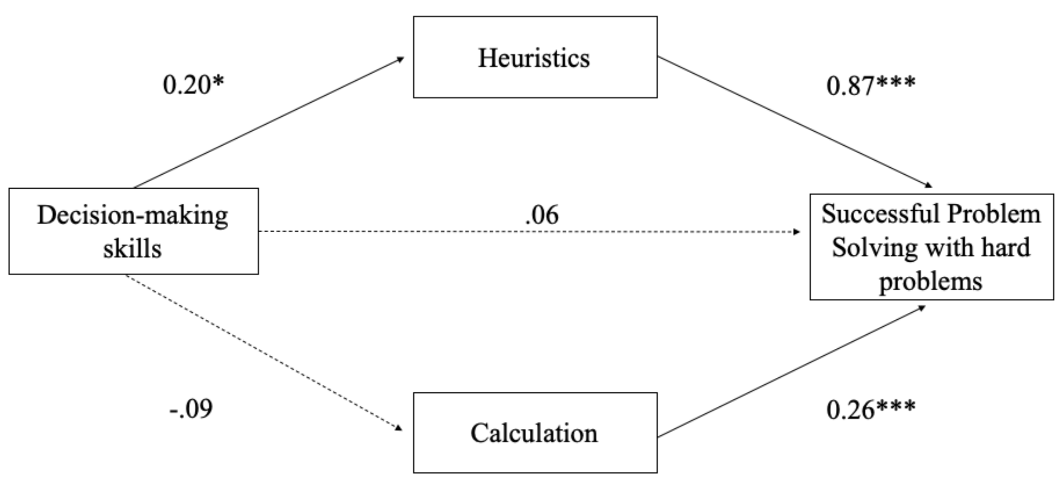
Did skilled decision makers use heuristics to solve hard problems?

[6] There was disagreement with less than 5% of the protocols between coders.

I constructed a structural cognitive path model using the lavaan package in R studio (Roussell, 2012) with accuracy as a dependent variable (i.e., accuracy on the three hard problems of the Berlin Numeracy Test). The independent variables were calculation use, heuristics use, and skilled decision making, which was operationalized as the total number correct on the Brief Risk Literacy Test. The model fit in Figure 2 was good: $\chi^2(1) = 0.07$, $p = 0.80$, CFI = 1.00, TLI = 1.03, SRMR = 0.01, RMSEA = 0.000 with 90% CI [.00, .20], such that skilled decision making predicted the use of heuristics ($\beta = .20$, $p < .05$), which in turn predicted successful hard problem solving ($\beta = .87$, $p < .05$). An analysis of indirect effects confirmed that heuristic use fully mediated the relationship between skilled decision-making and successful hard problem solving ($\beta = .172$, $SE = .075$, 95% CI [.024, .328]; Table 5). Skilled decision making was not related to the use of calculation strategies on hard problems.

Figure 2

Mediation model predicting successful problem solving of hard Berlin Numeracy Test problems



Note: Reported are standardized coefficients *$p < .05$, **$p < .01$, ***$p < .001$. Model estimated using lavaan package in R (Roussell, 2012).

Table 5

Mediation analysis summary

Total Effect (Decision Making Skill → Skilled Problem Solving)	Direct Effect (Decision Making Skill → Skilled Problem Solving)	Relationship	Indirect Effect	*SE*	Bootstrapped 95% CI	
					Low	Up
.208 (.014)	.060 (.084)	Decision making skill → **Heuristics →** Skilled problem solving	.172 (.015)	.075	.024	.328

Note: Effects were estimated with 1,000 bootstrap samples. SE represents standard error and CI represents confidence interval.

CHAPTER 4: DISCUSSION

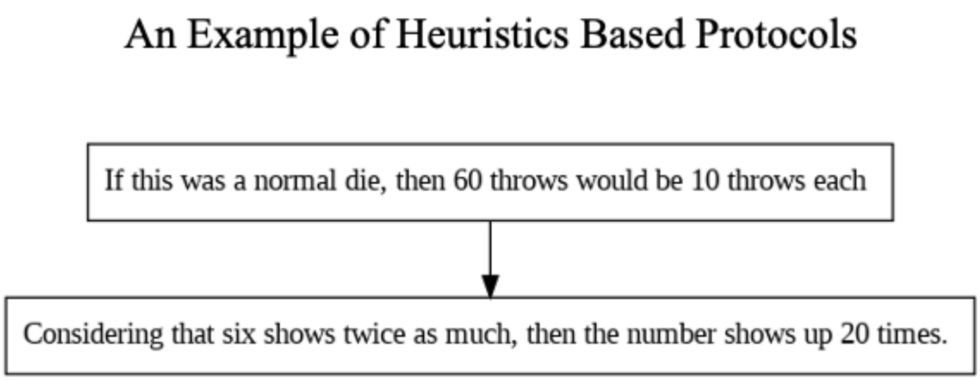
Protocol analysis indicates that when people are presented with a relatively easy problem on the Berlin Numeracy Test, they may generally rely on formal calculations for problem solving. However, when presented with more difficult problems from the Berlin Numeracy Test, people switch to heuristic-based strategies such that the relationship between skilled decision making and problem solving is fully explained by heuristic use. To the extent these findings generalize, it appears that skilled decision makers may often rely on heuristics to understand and solve many types of hard problems, rather than primarily relying on formal logic or calculation-based reasoning processes.

Theoretically, the observed pattern of results may largely be explained by two factors. One explanatory factor is that people typically need to conceptually understand a problem before they can determine what formulas they should use to calculate a solution (e.g., *comprehension before derivation*; Schoenfeld, 1979; 1989; 2022). This emphasis on the crucial role of representative understanding is consistent with research using models of language comprehension to explain effective mathematical problem solving (e.g., understanding is first guided by contextual and linguistic cues rather than formal reasoning; see Construction-Integration Theory; Kintsch, 1988; Kintsch et al., 1993; see also Cokely et al., 2018, in press; Feltz & Cokely, 2024a, 2024b). The other explanatory factor is that people's conceptual understanding can simplify the task's mathematical demands, as was the case for all three difficult problems on the Berlin Numeracy Test (e.g., *clarity simplifies complexity*; Polya, 1945; Cai et al., 2024). The difference in heuristics and calculation protocols on the 6-sided die question provides a clear illustration (e.g., 2-3 steps for heuristics v. 8 or more for calculation; see Figure 3). This benefit of conceptual clarity may also extend to some decision making tasks, potentially helping to explain how some fast and frugal heuristics outperform formal optimization processes (e.g., accurately identifying and ordering cue-validities for one-reason heuristic decision making; Gigerenzer, 2000, 2004; but see also Ericsson, 2003; Garcia-Retamero et al., 2019; Reyna & Brainerd, 2020; Reyna & Brust-Renck, 2023; Slovic, 2004, 2005).

Figure 3

Six sided-die example comparing heuristic based protocols to calculation based protocols

Question from Cokely et al. (2012): Imagine we are throwing a loaded die (6 sides). The probability that the die shows a 6 is twice as high as the probability of each of the other numbers. On average, out of these 70 throws how many times would the die show the number 6?



An Example of Calculation Based Protocols

In this scenario, if x represents a side of a normal die.

↓

Since all other sides are normal then they are represented as x"s, then it is x+x+x+x+x cause of the 5 normal sides, which is 5x

↓

Therefore, all possible outcomes have to equal the total amount of outcomes possible.

↓

Thus, you have 5x (representing the five sides) plus 2x (representing the number 6) equals to 70 (total amount of throws)

↓

Since, you add up like terms, then you get 7x=70, then you would divide each side by 7, which then gets you x=10

↓

Then, you plug x back into original equation which gets 5x equals to 50 and 2x equals to 20

↓

which matches the overall 70, which validates the equation

↓

Since, we are looking for how many times the number 6 appears, then that number is represented by 2x, which is 20.

Note: Both problem spaces display protocols that can be either heuristics or calculation. Categorization of response method was done based on the predominant strategy indicated by the protocols.

Limitations and Conclusion

Like all research, the current study has several limitations (e.g., use of a convenience sample; limitations of non-experimental protocol analysis investigation). The most noteworthy limitation may be the fact that I only conducted the protocol analysis on the Berlin Numeracy Test, which only consists of four numeracy test questions. Because of this, it is unclear whether the influence of heuristics may generalize to other difficult statistical numeracy problems. I speculate these findings will likely generalize to some extent because the first explanatory factor suggests it would be hard to derive correct equations without first developing an accurate conceptual understanding (i.e., comprehension before derivation). However, I think the second explanatory factor (i.e., clarity simplifies complexity) would be much less likely to generalize across the various kinds of difficult numeracy problems because complex calculations are sometimes a necessary part of mathematical problem solving (e.g., problems can be designed so there is no short cut).

It is noteworthy that the Berlin Numeracy Test is one of the strongest general predictors of decision making quality in the literature. This suggests the problems on the Berlin Numeracy Test are likely representative of some essential underlying skills and cognitive processes. It also seems plausible that the Berlin Numeracy Test is such a strong predictor in general because it specifically tests the ability to develop a representative understanding of a problem, rather than solely testing the ability to logically reason or to apply advanced procedural math knowledge (see Skilled Decision Theory, Cokely et al., 2018). Clearly more research is needed, but to the extent these findings generalize they seem likely to have many implications for the design of cognitive assessments (e.g., identifying problems that assess conceptual understanding along with other skills) and for the development of effective skill training programs (e.g., training for meaningful comprehension and metacognitive skills, before focusing on abstract logical reasoning).

REFERENCES

Allan, J. N., Cokely, E. T., Garcia-Retamero, R., Jenkins-Smith, H. C., Feltz, A., Gronlund, S. D., Ripberger, J. T., & Terry, R. (2021). Statistical numeracy norms and decision vulnerability benchmarks: A norm-referenced method for estimating the risk literacy difficulty level of choices and risk communications. https://shareok.org/handle/11244/330238

Ambrogio, S., Narayanan, P., Okazaki, A., Fasoli, A., Mackin, C., Hosokawa, K., Nomura, A., Yasuda, T., Chen, A., Friz, A., Ishii, M., Luquin, J., Kohda, Y., Saulnier, N., Brew, K., Choi, S., Ok, I., Philip, T., Chan, V., … Burr, G. W. (2023). An analog-AI chip for energy-efficient speech recognition and transcription. *Nature, 620*(7975), 768–775. https://doi.org/10.1038/s41586-023-06337-5

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

Cai, J., Koichu, B., Rott, B., & Jiang, C. (2024). Advances in research on mathematical problem posing: Focus on task variables. *The Journal of Mathematical Behavior*, 76, 101186. https://doi.org/10.1016/j.jmathb.2024.101186

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and Representation of Physics Problems by Experts and Novices*. *Cognitive Science*, *5*(2), 121–152. https://doi.org/10.1207/s15516709cog0502_2

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 1, pp. 7–76). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cho, J., Cokely, E. T., Feltz, A., & Garcia-Retamero, R. (under review). A 30-minute Tutorial on Icon Array Improves Risk Literacy and Decision Making Skills.

Cokely, E. T., Cho, J., Allan, J. N., Ramasubramanian, M., Nguyen, L.V., Ghazal, S., Perrin, O., Feltz, A., & Garcia-Retamero, R. (in press). Why Risk Literacy Matters: Knowledge is Power. In M. Williams, A. Kozbelt, F. Preckel, & R. Hoffman (Eds.), *The Cambridge Handbook of Expertise and Expert Performance* (3rd Edition). Cambridge University Press.

Cokely, E. T., Feltz, A., Ghazal, S., Allan, J. N., Petrova, D., & Garcia-Retamero, R. (2018). Skilled Decision Theory: From Intelligence to Numeracy and Expertise. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *The Cambridge Handbook of Expertise and Expert Performance* (2nd ed., pp. 476–505). Cambridge University Press. https://doi.org/10.1017/9781316480748.026

Cokely, E. T., & Kelley, C. M. (2009). Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation. *Judgment and Decision Making, 4*(1), 20–33. https://doi.org/10.1017/S193029750000067X

Cokely, E. T., Ghazal, S., Galesic, M., Garcia-Retamero, R., & Schulz, E. (2012). How to Measure Risk Comprehension in Educated Samples. In R. Garcia-Retamero & M. Galesic (Eds.), *Transparent Communication of Health Risks* (pp. 29–52). Springer New York. https://doi.org/10.1007/978-1-4614-4358-2_3

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring Risk Literacy: The Berlin Numeracy Test. *Judgment and Decision Making*, *7*(1), 25–47. https://doi.org/10.1017/S1930297500001819

De Groot, A. D. (1946). *Het denken van den schaker* [The thought of the chess player]. Noord-Hollandsche Uitgevers Maatschappij.

De Neys, W. (2006). Automatic–Heuristic and Executive–Analytic Processing during Reasoning: Chronometric and Dual-Task Considerations. *Quarterly Journal of Experimental Psychology*, *59*(6), 1070-1100. https://doi.org/10.1080/02724980543000123 (Original work published 2006)

Ericsson, K. A. (2003). The acquisition of expert performance as problem solving: Construction and modification of mediating mechanisms through deliberate practice. In J. E. Davidson & R. J. Sternberg (Eds.), *The psychology of problem solving* (pp. 31–83). Cambridge University Press. https://doi.org/10.1017/CBO9780511615771.003

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*(2), 211–245. https://doi.org/10.1037/0033-295X.102.2.211

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363–406. https://doi.org/10.1037/0033-295X.100.3.363

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

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). Cambridge, MA: Bradford Books/ MIT Press.

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. (1998). How to Study Thinking in Everyday Life: Contrasting Think-Aloud Protocols With Descriptions and Explanations of Thinking. *Mind, Culture, and Activity*, *5*(3), 178–186. https://doi.org/10.1207/s15327884mca0503_3

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring Numeracy without a Math Test: Development of the Subjective Numeracy Scale. *Medical Decision Making, 27*(5), 672–680. https://doi.org/10.1177/0272989X07304449

Feltz, A., & Cokely, E. T. (2024a). *Diversity and Disagreement: From Fundamental Biases to Ethical Interactions* (p. 289). Springer Nature.

Feltz, A., & Cokely, E. T. (2024b). Ethical interaction theory. In *Diversity and disagreement: From fundamental biases to ethical interactions* (pp. 227-243). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-61935-9

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

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Friederichs, H., Birkenstein, R., Becker, J. C., Marschall, B., & Weissenstein, A. (2020). Risk literacy assessment of general practitioners and medical students using the Berlin Numeracy Test. *BMC Family Practice*, *21*(1), 143. https://doi.org/10.1186/s12875-020-01214-w

Fuchs, L. S., Powell, S. R., Cirino, P. T., Schumacher, R. F., Marrin, S., Hamlett, C. L., Fuchs, D., Compton, D. L., & Changas, P. C. (2014). Does Calculation or Word-Problem Instruction Provide A Stronger Route to Pre-Algebraic Knowledge?. *Journal of educational psychology*, *106*(4), 990–1006. https://doi.org/10.1037/a0036793

Garcia-Retamero, R., Sobkow, A., Petrova, D., Garrido, D., & Traczyk, J. (2019). Numeracy and Risk Literacy: What Have We Learned so Far? *The Spanish Journal of Psychology*, 22, E10. doi:10.1017/sjp.2019.16

Garcia-Retamero, R., & Cokely, E. T. (2017). Designing Visual Aids That Promote Risk Literacy: A Systematic Review of Health Research and Evidence-Based Design Heuristics. *Human factors*, *59*(4), 582–627. https://doi.org/10.1177/0018720817690634

Garcia-Retamero, R., Cokely, E. T., Ghazal, S., & Joeris, A. (2016). Measuring Graph Literacy without a Test: A Brief Subjective Assessment. *Medical Decision Making*, 36(7), 854–867. https://doi.org/10.1177/0272989X16655334

Ghazal, S., Cokely, E. T., & Garcia-Retamero, R. (2014). Predicting biases in very highly educated samples: Numeracy and metacognition. *Judgment and Decision Making*, 9(1), 15–34. https://doi.org/10.1017/S1930297500004952

Gigerenzer, G. (2004). Fast and frugal heuristics: The tools of bounded rationality. In D. J. Koehler & N. Harvey (Eds.), *Blackwell handbook of judgment and decision making* (pp. 62–88). Blackwell Publishing. https://doi.org/10.1002/9780470752937.ch4

Gigerenzer, G. (2008). Why Heuristics Work. *Perspectives on Psychological Science*, *3*(1), 20-29. https://doi.org/10.1111/j.1745-6916.2008.00058.x

Hegarty, M., Mayer, R. E., & Monk, C. A. (1995). Comprehension of arithmetic word problems: A comparison of successful and unsuccessful problem solvers. *Journal of Educational Psychology*, *87*(1), 18–32. https://doi.org/10.1037/0022-0663.87.1.18

Kahneman, D. (2011). Thinking, fast and slow. Farrar, Straus and Giroux.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, *95*(2), 163–182. https://doi.org/10.1037/0033-295X.95.2.163

Kuhn, D. (2000). Metacognitive Development. *Current Directions in Psychological Science*, *9*(5), 178-181. https://doi.org/10.1111/1467-8721.00088

Mayer, R. E., & Hegarty, M. (1996). The process of understanding mathematical problems. In R. J. Sternberg & T. Ben-Zeev (Eds.), *The nature of mathematical thinking* (pp. 29–53). Lawrence Erlbaum Associates, Inc.

Mayer, R. E., & Hegarty, M. (2012). The process of understanding mathematical problems. In R. E. Mayer (Ed.), The nature of mathematical thinking (pp. 29–53). Routledge. https://doi.org/10.4324/9780203053270-4

Newell, A., & Simon, H. A. (1972). *Human problem solving.* Prentice-Hall.

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

Okan, Y., Janssen, E., Galesic, M., & Waters, E. A. (2019). Using the Short Graph Literacy Scale to Predict Precursors of Health Behavior Change. *Medical decision making : an international journal of the Society for Medical Decision Making*, *39*(3), 183–195. https://doi.org/10.1177/0272989X19829728

OpenAI. (2022). Whisper: Speech recognition model (Version 1.0) [Computer software]. GitHub. https://github.com/openai/whisper

OpenAI. (2023). ChatGPT (Mar 14 version) [Large language model]. https://chat.openai.com/chat

Payne, J. W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. *Organizational Behavior and Human Performance*, *16*(2), 366–387. https://doi.org/10.1016/0030-5073(76)90022-2

Peters, E., Slovic, P., Västfjäll, D., & Mertz, C. K. (2008). Intuitive numbers guide decisions. *Judgment and Decision Making*, *3*(8), 619–635. https://doi.org/10.1017/S1930297500001571

Peters, E. (2012). Beyond Comprehension: The Role of Numeracy in Judgments and Decisions. *Current Directions in Psychological Science, 21*(1), 31–35. https://doi.org/10.1177/0963721411429960

Peters, E. (2020). *Innumeracy in the Wild: Misunderstanding and Misusing Numbers*. Oxford University Press. https://doi.org/10.1093/oso/9780190861094.001.0001

Peters, E., McCaul, K. D., Stefanek, M., & Nelson, W. (2006). A heuristics approach to understanding cancer risk perception: Contributions from judgment and decision-making research. *Annals of Behavioral Medicine*, *31*(1), 45–52. https://doi.org/10.1207/s15324796abm3101_8

Point, S., & Baruch, Y. (2023). (Re)thinking transcription strategies: Current challenges and future research directions. *Scandinavian Journal of Management*, *39*(2), 101272. https://doi.org/10.1016/j.scaman.2023.101272

Pólya, G. (1945). How to solve it. Princeton University Press.

Pólya, G. (1954). Mathematics and plausible reasoning: Volume I: Induction and analogy in mathematics. Princeton University Press.

Python Software Foundation. (2023). Python (Version 3.x) [Computer software]. https://www.python.org

Qualtrics. (2023). Qualtrics Experience Management Platform [Computer software]. Qualtrics. https://www.qualtrics.com

Reyna, V. F. (2004). How People Make Decisions That Involve Risk: A Dual-Processes Approach. *Current Directions in Psychological Science*, *13*(2), 60-66. https://doi.org/10.1111/j.0963-7214.2004.00275.x (Original work published 2004)

Reyna, V. F. (2008). A Theory of Medical Decision Making and Health: Fuzzy Trace Theory. *Medical Decision Making*, *28*(6), 850–865. https://doi.org/10.1177/0272989X08327066

Reyna, V. F., & Brainerd, C. J. (2023). Numeracy, gist, literal thinking and the value of nothing in decision making. *Nature Reviews Psychology*, *2*(7), 421–439. https://doi.org/10.1038/s44159-023-00188-7

Reyna, V. F., & Brust-Renck, P. G. (2020). How representations of number and numeracy predict decision paradoxes: A fuzzy-trace theory approach. *Journal of Behavioral Decision Making*, *33*(5), 606–628. https://doi.org/10.1002/bdm.2179

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Schoenfeld, A. H. (2022). Why Are Learning and Teaching Mathematics So Difficult? In M. Danesi (Ed.), *Handbook of Cognitive Mathematics* (pp. 763–797). Springer International Publishing. https://doi.org/10.1007/978-3-031-03945-4_10

Schoenfeld, A. H. (1979). Explicit heuristic training as a variable in problem-solving performance. *Journal for Research in Mathematics Education*, *10*(3), 173–187. https://doi.org/10.5951/jresematheduc.10.3.0173

Schoenfeld, A. H. (1989). Problem solving in context(s). In R. I. Charles & E. A. Silver (Eds.), *The teaching and assessing of mathematical problem solving* (pp. 82–92). National Council of Teachers of Mathematics.

Schirrmeister, E., Göhring, A., & Warnke, P. (2020). Psychological biases and heuristics in the context of foresight and scenario processes. *Futures & Foresight Science*, *2*(2), e31. https://doi.org/10.1002/ffo2.31

Schwartz, L. M., Woloshin, S., Black, W. C., & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of internal medicine*, *127*(11), 966–972. https://doi.org/10.7326/0003-4819-127-11-199712010-00003

Skagerlund, K., Lind, T., Strömbäck, C., Tinghög, G., & Västfjäll, D. (2018). Financial literacy and the role of numeracy—How individuals' attitude and affinity with numbers influence financial literacy. *Journal of Behavioral and Experimental Economics, 74,* 18–25. https://doi.org/10.1016/j.socec.2018.03.004

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2004). Risk as analysis and risk as feelings: Some thoughts about affect, reason, risk, and rationality. *Risk Analysis*, *24*(2), 311–322. https://doi.org/10.1111/j.0272-4332.2004.00433.x

Slovic, P., Peters, E., Finucane, M. L., & Macgregor, D. G. (2005). Affect, risk, and decision making. *Health psychology : official journal of the Division of Health Psychology, American Psychological Association*, *24*(4S), S35–S40. https://doi.org/10.1037/0278-6133.24.4.S35

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General*, *127*(2), 161–188. https://doi.org/10.1037/0096-3445.127.2.161

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645–665. https://doi.org/10.1017/S0140525X00003435

Stanovich, K. E. (2009). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory? In J. S. B. T. Evans & K. Frankish (Eds.), *In two minds: Dual processes and beyond* (pp. 55–88). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199230167.003.0003

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645–665. https://doi.org/10.1017/S0140525X00003435

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, *94*(4), 672–695. https://doi.org/10.1037/0022-3514.94.4.672

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Rational thinking and cognitive sophistication: Development, cognitive abilities, and thinking dispositions. *Developmental Psychology*, *50*(4), 1037–1048. https://doi.org/10.1037/a0034910

APPENDIX

Materials for training "think out loud" method

Think Out Loud Training[7]

Think Out Loud Training[8]
Thank you for agreeing to take part in this study. I will start by familiarizing you with the procedure for "thinking out-loud". We are interested in knowing your thoughts as you come up with answers to the problems in this experiment. In order to do this, I am going to ask you think out loud as you work on the answer to some practice questions. What I mean by "think out-loud" is that I want you to say your thoughts out loud from the moment you finish hearing a practice question until you say the final answer. I would like you to talk out loud as much as you "comfortably" can during that time. Don't explain or summarize what you say, it does not even need to make sense. Just think out loud as if you are alone and speaking to yourself. Let's start with an easy question: Remember, don't just the instructions, but think out loud as you are solving the question. "What is the first letter after A?" (participant responds) B is correct. When I asked you "what is the first letter after A", you might have thought of B, but felt unsure of whether some thought occurred in between. In that case, you'd only say B (like you did). Only report thoughts that you have. Don't try to report unclear thoughts or thoughts you think you should have had or might have had, just think out loud as much as you can. Let's try something harder, please remember to "think out loud": "What is the sixth letter after B?" (participant responds) Good. Chances are that the letter "H" didn't immediately occur to you after hearing the question. You probably had to go through several steps to find the answer. Had you summarized your thinking during the last question, which we don't want, you might have said "I am finding the letter H by counting through the alphabet." But, when people solve this problem out loud, they usually say a sequence of individual letters such as B, then C, D, E, F and G, before they answer H. Remember, your never need to explain your thoughts just say exactly what you're thinking "out loud". Let's do another practice question:

[7] From Cokely et al. (in press)
[8] Every line represents a different slide presented to the participant. The participant answers the question under where "(participant responds)" is written. That part was not written in the training. Only the participants assigned to the verbal condition did the training. Participants in the silent condition did not do the training.

Please think aloud while you generate the answer. Are you ready. If you are, please click the arrow below. I’m going to show a dot grid and ask you to tell me how many dots there are in the grid. Please remember to think out loud as you solve the dot grid.

“How many dots are in this grid?” (participant responds)

• • • • • • • • •
• • • • • • • • •
• • • • • • • • •

Let’s try another grid. Just a reminder to avoid explaining or summarizing your responses. Just verbalize your thoughts as they comfortably come into your head. Ready? If you are, please click the arrow below.

“How many dots are in this grid?” (Participant responds)

• •

Now let’s try a different question to get you comfortable, as the thoughts come to mind. Ready? If you are, please click the arrow below

“How many dots are in this grid?” (Participant responds)

• • • • • •
• • • • • •
• • • • • •

Now you will see some additional math problems. Please verbalize your thoughts out loud as you solve them.

Let’s try an easy one first. Ready? If you are, please click the arrow below.
“What is 2+6?” (participant responds)

Let’s try something a little harder. Ready? If you are, please click the arrow below.
“What is 7x3?” (participant responds)

Let’s try one more problem like this. Ready? If you are, please click the arrow below.
“What is 8÷2?” (participant responds)

Let’s try something a little bit harder. Ready? If you are, please click the arrow below.
“What is 15x6?” (participant responds)

Let's try one more problem like this. Ready? If you are, please click the arrow below. "What is 70÷5?" (participant responds) Please describe the difference between thinking out loud and summarizing your thoughts (Participant responds) Would you like to continue with addition practice problems? (Participant responds with "yes" or "no" multiple choice)[9] "Please count the windows in your parents' house" (or in the house you grew up in).
Supplemental Think Out Loud Practice Materials[10]
- How many months begin with the letter J? - Would you like to continue with additional practice? (participant chooses "yes" or "no") - What is the fifth letter before M? - Would you like to continue with additional practice? (participant chooses "yes" or "no") - Name 5 animals in the zoo. - Would you like to continue with additional practice? (participant chooses "yes" or "no") - What is 14x10? - Would you like to continue with additional practice? (participant chooses "yes" or "no") - What is 13 x 7? - Would you like to continue with additional practice? (participant chooses "yes" or "no") - Name three different team sports.

[9] Participant had a choice to do extra practice. If yes was chosen, then participant went to the supplementary questions.

[10] After each supplementary question, the participant could choose whether to end the supplemental practice and return to the main training or continue on with the practice until the end of the supplementary question. The end of the supplementary question always went back to the final question of the training.

English as first language and effect on results[11]

T-test analyses between English first language and English non-first language (Verbal condition)

	English Speakers			Non-English Speakers			t-score	df	p-value	*d*
	N	*M*	*SE*	*N*	*M*	*SE*				
Berlin Numeracy Test (BNT)	88	1.00	0.09	14	1.14	0.29	0.57	100	0.57	0.16
Easy Problem BNT	88	0.57	0.05	14	0.57	0.14	0.02	100	0.98	<.01
Hard Problems BNT	88	0.43	0.07	14	0.57	0.20	0.76	100	0.45	0.22
Decision-Making Skill	88	1.97	0.10	14	1.86	0.31	-0.38	100	0.70	-0.11

T-test analyses between English first language and English non-first language (Silent condition)

	English Speakers			Non-English Speakers			t-score	df	p-value	*d*
	N	*M*	*SE*	*N*	*M*	*SE*				
Berlin Numeracy Test (BNT)	51	1.20	0.15	7	1.14	0.34	-0.12	56	0.90	-0.05
Easy Problem BNT	51	0.59	0.07	7	0.57	0.20	-0.08	56	0.93	-0.03
Hard Problems BNT	51	0.61	0.11	7	0.57	0.20	-0.12	56	0.91	-0.05
Decision-Making Skill	51	2.04	0.17	7	1.86	0.52	-0.67	56	0.51	-0.27

[11] Was not included in final analysis, most participants spoke English as their first language. Not enough non-first language English speaker to justify analysis.

Evaluating whether education level affected outcomes[12]

Regression ANOVA with education level predicting Berlin Numeracy Test Scores

	Sum of Squares	df	Mean Square	F	p-value
Regression	2.93	4	0.98	1.10	0.35
Residual	139.01	155	0.89		
Total	141.94	159			

Regression ANOVA with education level predicting easy problem accuracy on Berlin Numeracy Test

	Sum of Squares	df	Mean Square	F	p-value
Regression	0.53	3	0.18	0.71	0.55
Residual	38.57	156	0.25		
Total	39.10	159			

Regression ANOVA with education level predicting hard problem accuracy on Berlin Numeracy Test

	Sum of Squares	df	Mean Square	F	p-value
Regression	1.05	3	0.35	0.75	0.52
Residual	72.94	156	0.47		
Total	73.99	159			

Regression ANOVA with education level predicting hard problem accuracy on Berlin Numeracy Test

	Sum of Squares	df	Mean Square	F	p-value
Regression	1.05	3	0.35	0.75	0.52
Residual	72.94	156	0.47		
Total	73.99	159			

Regression ANOVA with education level predicting problem accuracy on decision-making skill problems

	Sum of Squares	df	Mean Square	F	p-value
Regression	1.53	3	0.51	0.44	0.72
Residual	179.32	156	1.15		
Total	180.84	159			

[12] Was not included in final analysis, most people in study either completed high school degree as highest education or some college. Other categories were not substantial enough for analysis.

Math courses completed affecting outcomes related to Berlin Numeracy Test and decision-making skill (Verbal condition)[13]

Regression whether the amount of math classes taken predicted accuracy on Berlin Numeracy Test

	Sum of Squares	df	Mean Square	F	p-value
Regression	4.02	1	4.02	5.59	0.02
Residual	71.94	100	0.72		
Total	75.96	101			

Regression whether the amount of math classes taken predicted accuracy on easy problem of Berlin Numeracy Test

	Sum of Squares	df	Mean Square	F	p-value
Regression	1.10	1	1.10	4.60	0.03
Residual	23.92	100	0.24		
Total	25.02	101			

Regression whether the amount of math classes taken predicted accuracy on hard problems of Berlin Numeracy Test

	Sum of Squares	df	Mean Square	F	p-value
Regression	0.92	1	0.92	2.27	0.14
Residual	40.34	100	0.40		
Total	41.25	101			

Regression whether the amount of math classes taken predicted correct problems of decision making skill

	Sum of Squares	df	Mean Square	F	p-value
Regression	3.54	1	3.54	3.72	0.06
Residual	95.21	100	0.92		
Total	98.75	101			

[13] Data was not used in final analysis as instructions were not clear to the participants what counted as college level math courses. Higher level math course people neglected to also choose the lower level math courses as well.

Math courses completed affecting outcomes related to Berlin Numeracy Test and decision-making skill (Silent condition)[14]

Regression whether the amount of math classes taken predicted accuracy on Berlin Numeracy Test

	Sum of Squares	df	Mean Square	F	p-value
Regression	1.75	1	1.75	1.55	0.22
Residual	63.16	100	1.13		
Total	64.91	101			

Regression whether the amount of math classes taken predicted accuracy on easy problem of Berlin Numeracy Test

	Sum of Squares	df	Mean Square	F	p-value
Regression	0.05	1	0.05	0.18	0.67
Residual	14.02	100	0.25		
Total	14.07	101			

Regression whether the amount of math classes taken predicted accuracy on hard problems of Berlin Numeracy Test

	Sum of Squares	df	Mean Square	F	p-value
Regression	1.23	1	1.23	2.25	0.14
Residual	30.65	100	0.55		
Total	31.88	101			

Regression whether the amount of math classes taken predicted correct problems of decision making skill

	Sum of Squares	df	Mean Square	F	p-value
Regression	0.46	1	0.46	0.32	0.58
Residual	81.54	100	1.46		
Total	82.00	101			

[14] Data was not used in final analysis as instructions were not clear to the participants what counted as college level math courses. Higher level math course people neglected to also choose the lower level math courses as well.

Berlin Numeracy Test Item Response Accuracy

Correct responses to each item on the Berlin Numeracy Test

Berlin Numeracy Test Problem by statistical concept	Correct Responses	Incorrect Responses	Proportion Correct
Problem 1 (Expected value)	63	39	0.62
Problem 2 (Conditional Probability)	25	77	0.25
Problem 3 (Weighted Probability)	14	88	0.14
Problem 4 (Inverse Conditional Probability)	7	95	0.07
Total	109	299	0.27

Note: Last cell in the last row represents which represents the average problem solving accuracy on the Berlin Numeracy Test.

Correct responses to each item on Berlin Numeracy Test by problem solving method

	Ambiguous		Calculation		Heuristic	
	Overall Responses (Correct Reponses)	Proportion	Overall Responses (Correct Reponses)	Proportion	Overall Responses (Correct Reponses)	Proportion
BNT1	11 (2)	0.18	83 (55)	0.66	8 (6)	0.75
BNT2	15 (0)	0	59 (1)	0.02	28 (24)	0.89
BNT3	13 (0)	0	74 (3)	0.04	15 (11)	0.73
BNT4	45 (2)	0.04	50 (1)	0.02	7 (4)	0.67
Total	84 (4)	0.05	266 (60)	0.23	58 (45)	0.78

Note: Total row represents the accuracy by problem solving method except for the last column which represents overall accuracy.

Intercorrelation tables

Intercorrelations for all problems on the Berlin Numeracy Test

	1	2	3	4	5
1. Problem Solving					
2. Heuristics	.72***				
3. Calculation	.56***	.04			
4. Ambiguous	.17	.18	-.21*		
5. Decision making skills	.29***	.18	.16	-.01	

Note: *p<.05, **p<.01, ***p<.001

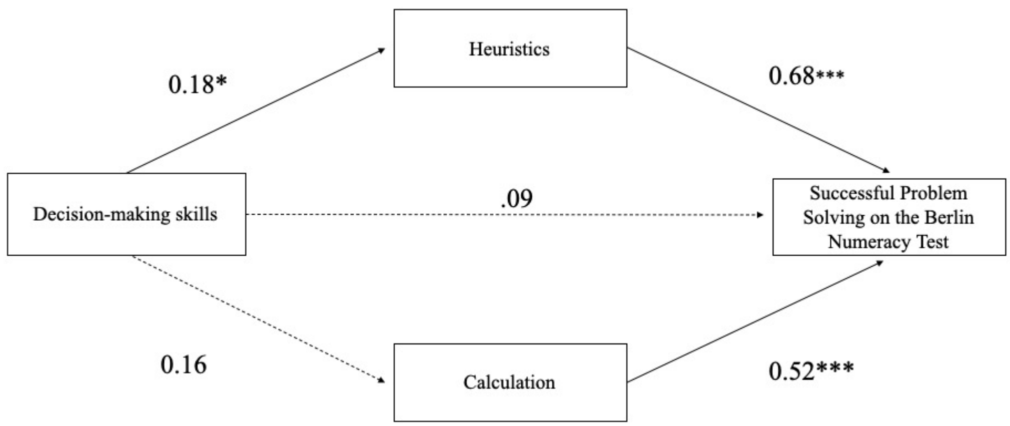
Intercorrelation for easy problems and hard problems on the Berlin Numeracy Test

	1	2	3	4	5	6	7
1. Problem Solving (Easy Problem)							
2. Heuristics (Easy Problem)	.13						
3. Calculation (Easy Problem)	.74***	-.27**					
4. Problem Solving (Hard Problems)	.16	.08	.16				
5. Heuristics (Hard Problems)	.13	-.02	.17	.88***			
6. Calculation (Hard Problems)	.01	-.06	.12	.26**	.01		
7. Decision-Making Skill Problems	.25**	-.02	.22*	.21*	.20*	-.09	

Note: *p<.05, **p<.01, ***p<.001

Mediation models[15]

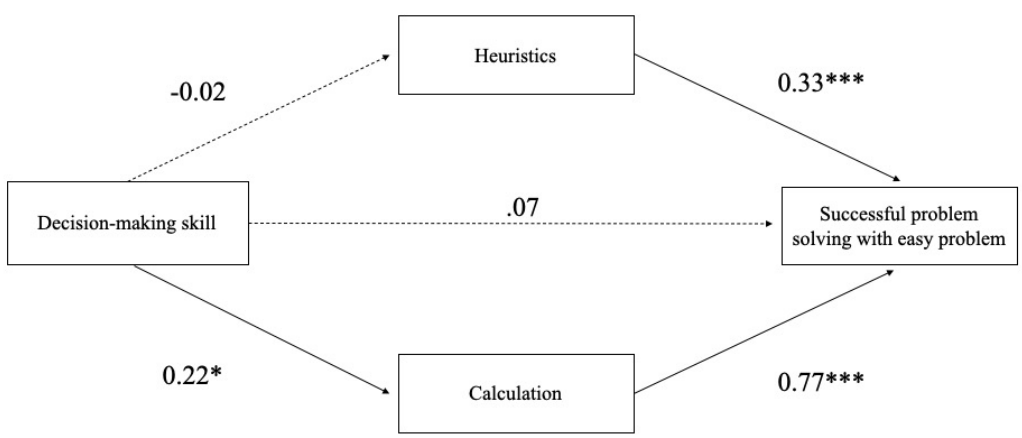
Mediation model predicting successful problem solving of the Berlin Numeracy Test (all problems)



$\chi^2(1)$= 0.02, p=0.886, CFI= 1.000, TLI= 1.036, SRMR= 0.006, RMSEA= 0.000 with 90% CI [.00, .13]

Note: Reported are standardized coefficients *p<.05, **p<.01, ***p<.001. Model estimated using *lavaan* package in R (Roussell, 2012).

Mediation model predicting successful problem solving of the easy BNT problem



$\chi^2(1)$= 7.93, p=0.005, CFI= 0.944, TLI= 0.633, SRMR= 0.124, RMSEA= 0.261 with 90% CI [.12, .42]

Note: Reported are standardized coefficients *p<.05, **p<.01, ***p<.001. Model estimated using *lavaan* package in R (Roussell, 2012).

[15] Excluded from final analysis.

Demographic questions set 1

Demographic questions asked to participants at the end of the study[16]

What sex were you assigned at birth, on your original birth certificate?

- Male
- Female
- Don't Know
- Prefer not to answer

How old are you?

- Under 18
- 18 - 24
- 25 - 34
- 35 - 44
- 45 - 54
- 55 - 64
- 65 - 74
- 75 - 84
- 85 or older
- Prefer not to say

Which of the following best describes your race?

- White
- Black or African American
- American Indian or Alaska Native
- Asian
- Native Hawaiian or Pacific Islander
- Other
- Prefer not to answer

What is your current marital status?

- Married
- Widowed
- Divorced

[16] This demographic survey was used for the first 40 of 102 participants

- Separated
- Never married
- Prefer not to answer

Do you consider yourself to be Hispanic, Latino, or Spanish or to have Hispanic, Latino, or Spanish origins?

- Yes
- No
- Prefer not to answer

Is English your first langauge

- Yes
- No
- Prefer not to answer

What is the highest level of education you have COMPLETED?

- Less than high school
- High school graduate / GED
- Some college
- 2 year degree
- 4 year degree
- Professional degree
- Doctorate
- Prefer not to answer

Which one of the following BEST describes your current employment status?

- Employed full time
- Employed part time
- Unemployed looking for work
- Unemployed not looking for work
- Retired
- Student
- Disabled
- Prefer not to answer

How many classes involving statistics have you done in your lifetime? (please give your best estimate)

How many math classes have you taken in college or equivalent (i.e. Community college, AP classes in high school, etc.)? This includes if you have tested out of the courses. Please take a look at the following math classes on this list and choose. You may choose more than one.

- MATH 1471. Mathematics for Critical Thinking Corequisite
- MATH 1473. Mathematics for Critical Thinking
- MATH 1503. College Algebra
- MATH 1523. Precalculus and Trigonometry
- MATH 1643. Functions and Modeling for Business, Life and Social Sciences
- MATH 1743. Calculus I for Business, Life and Social Sciences
- MATH 2123. Calculus II for Business, Life and Social Sciences
- MATH 2423. Calculus and Analytic Geometry II
- MATH 2433. Calculus and Analytic Geometry III.
- MATH 2443. Calculus and Analytic Geometry IV.
- MATH 2934. Differential and Integral Calculus III.

Demographic questions set 2

Demographic questions asked to participants at the end of the study[17]

What sex were you assigned at birth, on your original birth certificate?

- Male
- Female
- Don't Know
- Prefer not to answer

What is your current gender? [Mark only one]

- Male
- Female
- I use a different term
- Prefer not to answer

How old are you?

- Under 18
- 18 - 24
- 25 - 34
- 35 - 44
- 45 - 54
- 55 - 64
- 65 - 74
- 75 - 84
- 85 or older
- Prefer not to say

Which of the following best describes your race?

- White
- Black or African American
- American Indian or Alaska Native
- Asian
- Native Hawaiian or Pacific Islander
- Other

[17] This demographic survey was used for the last 62 out of 102 verbal participants

- Prefer not to answer

What is your current marital status?
- Married
- Widowed
- Divorced
- Separated
- Never married
- Prefer not to answer

Do you consider yourself to be Hispanic, Latino, or Spanish or to have Hispanic, Latino, or Spanish origins?
- Yes
- No
- Prefer not to answer

Is English your first langauge
- Yes
- No
- Prefer not to answer

What is the highest level of education you have COMPLETED?
- Less than high school
- High school graduate / GED
- Some college
- 2 year degree
- 4 year degree
- Professional degree
- Doctorate
- Prefer not to answer

Which one of the following BEST describes your current employment status?
- Employed full time
- Employed part time
- Unemployed looking for work
- Unemployed not looking for work

- Retired
- Student
- Disabled
- Prefer not to answer

How many classes involving statistics have you done in your lifetime? (please give your best estimate)

How many math classes have you taken in college or equivalent (i.e. Community college, AP classes in high school, etc.)? This includes if you have tested out of the courses. Please take a look at the following math classes on this list and choose. You may choose more than one.

- MATH 1471. Mathematics for Critical Thinking Corequisite
- MATH 1473. Mathematics for Critical Thinking
- MATH 1503. College Algebra
- MATH 1523. Precalculus and Trigonometry
- MATH 1643. Functions and Modeling for Business, Life and Social Sciences
- MATH 1743. Calculus I for Business, Life and Social Sciences
- MATH 2123. Calculus II for Business, Life and Social Sciences
- MATH 2423. Calculus and Analytic Geometry II
- MATH 2433. Calculus and Analytic Geometry III.
- MATH 2443. Calculus and Analytic Geometry IV.
- MATH 2934. Differential and Integral Calculus III.