Adaptive judgment in the cognitive reflection test: A computational analysis

Yijin Hu[1] & Sudeep Bhatia (bhatiasu@sas.upenn.edu)[1]
[1]Department of Psychology, University of Pennsylvania

Abstract

The Cognitive Reflection Test (CRT) is widely used in reasoning and decision-making research, yet it lacks a formal cognitive foundation. As a result, debates persist over why CRT scores correlate with mathematical ability, why individuals differ in performance, and which mental processes drive observed response patterns. We address these questions using an ecological perspective combined with computational cognitive modeling. We characterize the learning environment by assembling a large dataset of grade-school verbal math problems and specify a learning mechanism that maps linguistic features of problems to arithmetic operations based on prior experience. Our model reproduces the characteristic intuitive errors elicited by CRT problems and explains them as byproducts of adaptive cognition. It further generates novel predictions about how environmental structure and problem wording influence strategy selection and performance, which we test in two preregistered experiments. Overall, our work provides a principled, quantitative account of CRT performance grounded in adaptive generalization.

Keywords: cognitive reflection task; adaptive judgment; computational modeling; intuitive judgment

Introduction

The Cognitive Reflection Test (CRT) consists of short problems, mathematical in nature, that reliably elicit incorrect but compelling intuitive responses (Frederick, 2005). For example, in the classic bat and ball problem, participants are asked: "A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?" The common answer, which is the consequence of the intuitive subtraction operation, is 10 cents. By contrast, the correct answer, 5 cents, needs to be solved by algebraic analysis of a system of equations jointly describing the sum and difference of the prices of the bat and the ball.

Scores on the CRT correlate with a wide range of cognitive and behavioral outcomes, and the test is often interpreted as an individual difference measure indexing the ability to override intuitive responses in favor of reflection. However, this interpretation struggles to explain several empirical findings. CRT accuracy is strongly associated with numeracy and mathematical ability, constructs that are theoretically distinct from reflection or inhibition (Campitelli & Gerrans, 2014; Erceg et al., 2020; Sinayev & Peters, 2015; Welsh et al., 2013). Moreover, younger students often outperform older non-students, and men tend to outperform women, despite evidence that self-regulatory capacity and inhibitory control are not higher in these groups (Brañas-Garza et al., 2019). Cognitive process data further complicate the standard account. Many participants who ultimately answer correctly appear to generate the correct response seemingly intuitively, while some who answer incorrectly nonetheless reflect extensively on their initial answer (Bago & De Neys, 2019; Patel et al., 2019; Szaszi et al., 2017). Moreover, errors are common even in the absence of an intuitive lure (Erceg et al., 2020; Patel et al., 2019), and incentives, warnings, hints, and explicit instructions to reflect (manipulations that are typically assumed to promote mental effort) have little effect on CRT performance (Brañas-Garza et al., 2019). Finally, CRT scores do not correlate with self-reported use of intuition (Pennycook et al., 2016).

We propose an alternative framework in which CRT errors arise from adaptive generalization shaped by the learning environment. Drawing on ecological and strategy-selection perspectives (Brunswik, 1955; Dhami et al., 2004; Lieder & Griffiths, 2019; Marewski & Schooler, 2011), we argue that individuals learn associations between linguistic features of verbal math problems and the operations required to solve them. When novel problems resemble familiar problem types in wording or structure, previously successful strategies are applied, even when they are inappropriate, yielding CRT-style errors. We formalize this account using a large dataset of grade-school verbal math problems (Miao et al., 2020) and represent problem language using large language model embeddings. A simple classifier trained on these representations accurately predicts the correct operations for standard math problems but systematically reproduces human-like errors on CRT items. The model further predicts that CRT performance depends on both environmental exposure and problem wording. We confirm these predictions in two preregistered experiments by manipulating linguistic framing and recent mathematical experience.

Overview of Framework

We model performance on the Cognitive Reflection Test (CRT) as a strategy selection problem based on generalization from prior experience with verbal math problems. On each trial, the input is a novel problem statement and the output is a prediction about the mathematical operation required to solve it (e.g., subtraction or algebra). This framework requires three core components: an input representation that maps problem text onto features, a learning environment that captures prior experience with math problems and their correct solutions, and a learning rule that generalizes from this experience to new problems. Together, these components define a formal computational account of how linguistic cues guide operation selection in the CRT.

In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society*

We represent problem text using sentence-level embeddings derived from transformer-based language models, specifically SBERT embeddings that encode entire problem statements as fixed-length vectors optimized to preserve sentence-level semantic similarity (Reimers & Gurevych, 2019). These embeddings capture both word content and higher-order linguistic structure, including syntax and relational phrasing. As a comparison, we also use a bag-of-words approach based on Word2Vec embeddings, in which pre-trained word vectors are averaged across the words in a problem (Mikolov et al., 2013). While this approach captures thematic and lexical information, it discards word order and sentence structure, allowing us to assess whether such higher-level linguistic features are critical for strategy selection.

The learning environment is operationalized using a large dataset of grade-school verbal math problems annotated with their underlying operations (Miao et al., 2020), which serves as a proxy for educational experience. We train classifiers to map linguistic representations of problems onto operations using ridge-regularized linear models as our primary learning rule, with k-nearest neighbors as a robustness check. These models implement strategy selection as a similarity-based categorization process, closely related to feature-based and exemplar models of judgment and learning, enabling us to examine how experience, representation, and problem framing jointly shape CRT performance.

Learning Environment

Text	SBERT Embedding	Operation
Ellen has 6 more ...	[0.48, 0.13, ..., 0.21]	Subtraction
George has 5 blue ...	[0.82, 0.19, ..., 0.99]	Addition
14 red plums cost ...	[0.10, 0.87, ..., 0.02]	Multiplication
A cup has 3 ...	[0.14, 0.34, ..., 0.41]	Algebra
...	...	...
It takes a train ...	[0.26, 0.89, ..., 0.01]	Subtraction

Cognitive Reflection Test

Text	SBERT Embedding	Operation
A bat and a ball ...	[0.88, 0.07, ..., 0.51]	?
It takes 5 machines ...	[0.10, 0.33, ..., 0.42]	?
In a lake, there is ...	[0.56, 0.29, ..., 0.81]	?

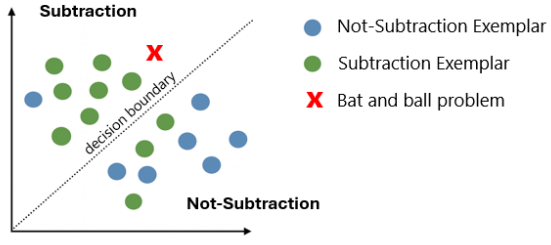

Figure 1: Our framework trains classifiers on language embeddings of grade-school math problems to predict the strategy evoked by the problem.

Study 1: Adaptive Judgment

Before turning to the CRT, we first tested whether our model could adaptively learn the mapping between problem wording and the correct mathematical operation. We focused on five operations commonly encountered in grade-school math, addition, subtraction, multiplication, division, and algebra. This step is essential: if the model cannot reliably identify the correct operation for standard problems, it cannot meaningfully explain CRT performance.

We evaluated model performance using the ASDiv dataset of grade-school verbal math problems, training separate binary (one-vs-rest) classifiers for each operation. To ensure consistency, we balanced each training set by pairing positive examples with an equal number of negatives. We compared two linguistic representations (SBERT sentence embeddings and Word2Vec bag-of-words embeddings) and two learning rules (ridge regression and k-nearest neighbors), using five-fold cross-validation to assess out-of-sample accuracy.

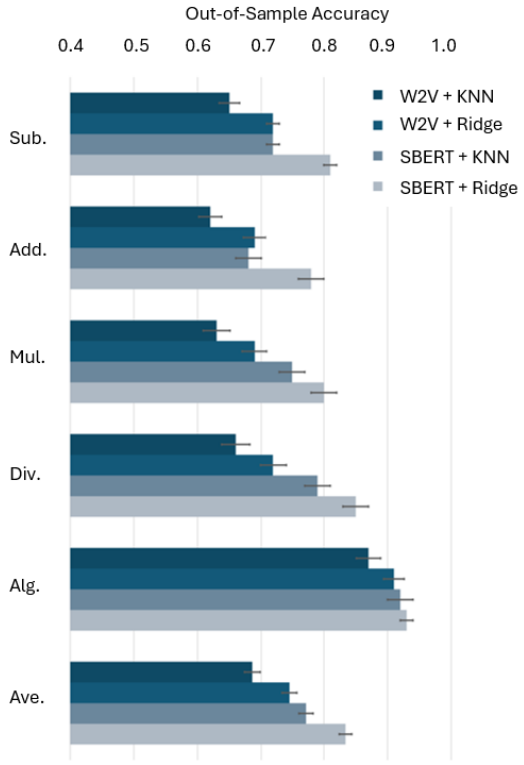

Figure 2: Our model successfully predicts the correct operation on the ASDiv dataset. Error bars indicate +/-1 SE.

All models performed significantly above chance, demonstrating that linguistic features reliably predict the required mathematical operation. The best-performing model was a ridge classifier trained on SBERT embeddings, which achieved high accuracy across all operations (average accuracy of 83%). Models using Word2Vec embeddings or k-nearest neighbors performed moderately well but consistently worse, indicating the importance of sentence-level structure and weighted feature integration. To test

generalizability, we evaluated the trained models on a new dataset of 2,500 LLM-generated grade-school math problems spanning the same five operations. Performance remained well above chance, with the SBERT-based ridge classifier again outperforming alternatives (average accuracy of 80%).

Finally, we improved performance by training classifiers on the full ASDiv dataset using class-weighted ridge regression rather than subsampling to balance classes. This approach increased accuracy across all operations (average accuracy of 88%), confirming that the model benefits from greater exposure to the learning environment. We use this class-weighted SBERT–ridge model as the primary implementation in all subsequent analyses.

Study 2: Errors on the CRT

Study 1 showed that our model reliably learns to map linguistic features of verbal math problems to the correct arithmetic operations. We next examined how this same model behaves when applied to CRT problems, which deliberately resemble familiar arithmetic items while requiring different underlying operations, making them a natural test case for our framework. We focus in particular on the bat-and-ball problem, where subtraction and algebra are in direct competition and both operations are well represented in the ASDiv dataset that is the foundation of our analysis.

Intuitive Operations on CRT Problems

We applied the best-performing model from Study 1 to the three original CRT problems. For the bat-and-ball problem, the model assigned very high probability to subtraction (96%) and nearly as high probability to algebra (92%), while assigning negligible probability to other operations. This mirrors human response patterns, in which many participants apply subtraction and give the canonical incorrect answer, while others apply algebra and respond correctly. A similar correspondence emerged for the remaining items. For example, for the widget problem, the model assigned probability 100% to division and 39% to multiplication, reflecting the common intuitive strategy of dividing by five and then scaling by 100. Likewise, for the lily-pad problem, multiplication received the highest probability (88%), consistent with the intuitive but incorrect tendency to halve 48. We found similar results for other intuitive judgment problems not on the CRT (results omitted due to space constraints).

To test for robustness, we applied our model to 100 GPT-generated variants of the bat and ball problem that preserved the logical structure of the original item while varying surface features. Across these variants, the model again assigned high probability to both subtraction (M = 65%) and algebra (M = 62%), with other operations receiving substantially lower probabilities.

Changing Wording

Although the model successfully identifies algebraic structure in standard grade-school problems, bat-and-ball items closely resemble familiar subtraction problems in both vocabulary and phrasing (e.g., references to objects, prices, and comparative expressions such as “more than”). These surface cues appear to activate subtraction despite the underlying algebraic structure required for the correct solution. To test this hypothesis, we rewrote each bat-and-ball variant into a decontextualized form that preserved the numerical relations while removing concrete objects, monetary framing, and comparative language. For example, an original item such as “A cap and a jacket cost $44 in total. The jacket costs $32 more than the cap. How much does the cap cost?” was rewritten as “The sum of two numbers is 44. The difference between the numbers is 32. What is the smaller number?”

This manipulation dramatically altered model behavior. The mean probability assigned to algebra on the 100 bat and ball variants increased from 62% to 92%, while the probability assigned to subtraction dropped from 65% to 57%. Moreover, the proportion of items for which algebra was favored over subtraction increased from 41% to 91% (all differences significant at $p < .01$). Because the learning algorithm and training data were held fixed, these shifts can only be attributed to changes in wording.

Changing the Learning Environment

While rewording manipulates cues at the time of judgment, another route to altering strategy selection is to change the learning environment itself. From an ecological perspective, the availability of a strategy depends on how frequently and recently it has been encountered. This provides a natural explanation for individual differences in CRT performance: individuals with greater exposure to algebra may be more likely to retrieve and apply algebraic strategies, even when problems are framed in misleading ways.

We tested this idea by systematically manipulating the training environment. First, we simulated curricular exposure by training our models on grade 1–5 problems (which largely exclude algebra) versus grade 1–6 problems (which include algebra). With this approach, we found that our models trained on grade 1–5 generated an average probability of subtraction of 74% on the 100 bat and ball variants. In contrast, when algebra examples were included in the training data (grades 1–6), the subtraction probability dropped to 65% ($p < .05$). Again, the algebra detector trained on the same grade 1–6 data also generated a nearly identical mean probability of 62% (since there were no algebra problems before grade 6, this classifier was not applied to the grade 1–5 training data).

To isolate the effect of relative exposure, we systematically varied the proportion of algebra versus subtraction problems in the training data by training separate classifiers with 25%, 50%, or 100% of the available problems from each category while holding the opposing category fixed as negatives. When applied to 100 LLM-generated bat-and-ball variants, increasing algebra exposure raised the mean probability of algebra classification from 9.4% to 31.5%, whereas increasing subtraction exposure raised the mean probability of subtraction classification from 25% to 63% (all ps < .01).

Together with the grade-level manipulation, these results show that the composition of the learning environment strongly and predictably shapes strategy selection on CRT problems.

Study 3: Empirical Test of Problem Wording

The results from Study 2 showed that rewording bat-and-ball-style problems significantly influenced our model's responses. To test whether humans showed a similar sensitivity to the wording of bat and ball problems, we conducted a behavioral experiment. We hypothesized that rewording CRT questions from contextualized to decontextualized formats would increase accuracy and reduce the likelihood of incorrect responses in our participants, supporting our theory of CRT errors.

Methods

We recruited 402 adult participants (age 18 or older) from Prolific. All participants were fluent in English, located in the United States, and had a 100% approval rate with 5–100 prior submissions on the platform. Participants provided informed consent before beginning the study. The target sample size was determined a priori based on a power analysis and results from a pilot study, which indicated that a sample of this size would provide 80% power to detect expected differences in accuracy between conditions at $\alpha = .05$. The study was preregistered.

The stimulus set consisted of 20 bat-and-ball-style problems—10 contextualized and 10 decontextualized—adapted from the GPT-generated items used in Study 2. Each pair of items had identical numerical values but differed in wording. Each participant was told they would be solving a math problem and instructed to provide a numerical response. They were presented with one bat-and-ball-style question, either in its original form or in a decontextualized version.

Results and Discussion

We first examined whether rewording CRT-style problems into a decontextualized format improved participants' performance. Accuracy was computed as the proportion of participants who gave the correct answer, and subtraction operation rate was defined as the proportion of participants giving the canonical error response, which occurs when participants misapply subtraction (e.g., answering 10 cents when the correct answer is 5 cents).

There were 203 participants in the original condition and 199 participants in the decontextualized condition. Participants were significantly more likely to answer correctly in the decontextualized condition than in the original condition with an accuracy rate of 80% vs 61% ($z = -4.13$, $p < .001$). Conversely, participants gave significantly fewer subtraction error responses in the decontextualized condition than in the original condition, with rate of 16% vs. 36% ($z = 4.44$, $p < .001$). (Figure 3) These results indicate that removing contextual framing improved accuracy while reducing subtraction operation responses, and align with our computational model and support the view that CRT errors emerge from overgeneralized patterns of language-operation mapping learned from the ecological structure of the environment.

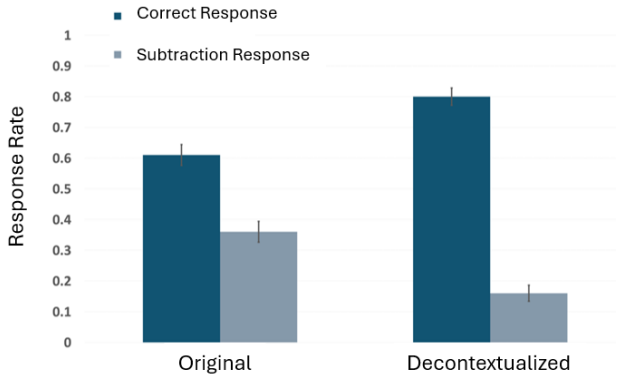

Figure 3: Participants are more accurate and less likely to use subtraction in the decontextualized condition compared to the original condition in Study 3. Error bars indicate +/- 1SE.

Study 4: Empirical Test of Experience

While Study 3 demonstrated that wording influenced accuracy on CRT problems, our results also suggested that prior experience plays a critical role. This raises an important question: can changing proximate experience reduce CRT errors? For example, if an individual is exposed to algebra immediately before attempting the bat and ball problem, will they be less likely to misapply subtraction? It is worth noting that our tests in Study 2 relied on training the model on a large dataset of algebra or subtraction problems, but proximate experience may operate even with only a small number of recent exposures. In particular, even though recency is not explicitly built into our model, it follows naturally from an ecological perspective in which the likelihood of applying a given strategy depends on its availability in memory and its similarity to current problems. Following this logic, the main goal of Study 4 was to test whether providing participants with a small sample of algebra problems immediately before the bat and ball task would shift their responses toward the correct solution.

Methods

We recruited 150 adult participants (age 18 or older) from Prolific. All participants were fluent in English, located in the United States, and had a 100% approval rate with 5–100 prior submissions on the platform. Participants provided informed consent before beginning the study. An a priori power analysis conducted in G*Power using pilot data indicated that a minimum of 36 participants would be sufficient to detect a medium effect with 80% power at $\alpha = .05$. The final sample size ($N = 150$) exceeded this minimum and was determined by available funding, providing increased precision in the estimates. The study was preregistered.

Participants were randomly assigned to one of two between-subjects conditions. In the algebra condition, participants completed two short algebra problems involving two unknown variables. In the subtraction condition, participants completed two simple math problems involving only subtraction. After completing the math problems, all participants were presented with one standard bat-and-ball-style question randomly drawn from a set of 10 versions adapted from Study 3. These items preserved the logic of the classic bat and ball problem but varied in objects and numerical values..

Results and Discussion

We analyzed participants' responses to the bat and ball problem as a function of the type of problems they completed beforehand (algebra vs. subtraction). As before, accuracy was coded as correct versus incorrect, and subtraction operation rate was defined as the proportion of participants who gave the canonical error response, which occurs when subtraction was misapplied. We conducted two complementary analyses. The unfiltered analysis included all participants with a valid condition label, regardless of how they performed on the prerequisite problems. The filtered analysis restricted the sample to participants who answered all algebra or subtraction items correctly, ensuring that these individuals were appropriately primed before attempting the bat and ball problem.

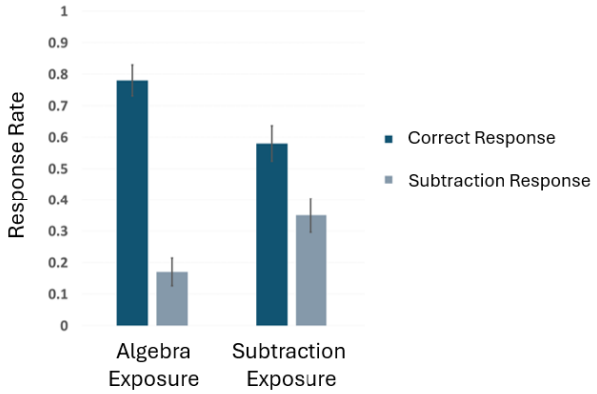

Figure 4: Participants are more accurate and less likely to use subtraction when previously exposed to algebra than when previously exposed to subtraction in Study 4.

In the unfiltered analysis, there were 72 participants in the algebra condition, and 78 participants in the subtraction condition. Participants in the algebra condition showed significantly higher accuracy of 78% than the 58% accuracy rate of those in the subtraction condition ($z = 2.62$, $p < .01$). Providing the subtraction error response was also less common in the algebra condition than in the subtraction condition (17% vs 35%) ($z = -2.50$, $p < .05$) (Figure 4). The filtered analysis yielded a similar pattern. After excluding participants who failed their prerequisite problems, 62 remained in the algebra condition and 75 in the subtraction condition. Accuracy was again higher in the algebra condition (84% vs 60%) ($z = 3.06$, $p < .01$), and subtraction operation rate was lower in the algebra condition (11% vs 35%) ($z = -3.19$, $p < .01$).

General Discussion

Cognitive rules are adapted to solve problems accurately and efficiently within the environments in which they are learned. Through repeated exposure, people acquire associations between problem features and solution strategies, which enable rapid, low-effort judgments that approximate optimal reasoning for the kinds of problems most frequently encountered. However, because these rules are tuned to the statistical structure of past experience, they can also generate systematic errors when applied outside their natural domain. In verbal math and reasoning problems, if the linguistic cues of a novel problem resemble those of a familiar problem class, individuals may overgeneralize from prior experience, leading to intuitively compelling but incorrect answers.

The Cognitive Reflection Test (CRT) provides a clear illustration of this phenomenon. Its problems are intentionally phrased to evoke simple arithmetic operations like subtraction, even though their correct solutions require more complex or abstract operations. Although researchers have long treated the CRT as a measure of intuitive response tendency and linked it to a wide variety of individual difference measures like impatience and numeracy, relatively little work has explicitly analyzed the cognitive mechanisms that elicit these intuitive responses in the first place. This omission is particularly striking given that psychology already possesses well-established theories of high-level cognition (and, in particular, mathematical cognition and arithmetic problem solving) that directly address how people learn and apply strategies in such tasks.

Ignoring prior theory in favor of treating the CRT purely as a measure of intuition limits our ability to predict performance differences across individuals and contexts. Why is it that men tend to outperform women or that high school students often score higher than adults? These patterns cannot plausibly be attributed to a simple ability to suppress impulsive responses. Likewise, the strong correlations between numeracy and CRT accuracy are not fully explained by an intuition or inhibition-based account; if that were the case, we would expect numeracy to correlate equally with other tasks requiring impulse control but unrelated to mathematical reasoning. Process data on the CRT present us with another set of puzzles: even among individuals who arrive at the correct answer, many appear to do so effortlessly while others who respond incorrectly nonetheless engage in extensive deliberation. Errors also occur even when no intuitive lure is present and appear to be insensitive to incentives, hints, warnings, and instructions to engage in reflection. Together, these findings suggest that the CRT is not merely a measure of one's capacity to override intuition. Rather, responses on the CRT depend on the accessibility of appropriate mathematical operations and representational structures, which in turn reflect differences in the learning

environment or contextual factors like problem wording and proximate experience.

This paper addresses these challenges by integrating language modeling, cognitive theory, and behavioral experimentation. We demonstrate that a simple strategy learning framework trained on grade school math problems not only performs well on those same problems but also reproduces the characteristic response errors observed in human CRT performance. The same framework accurately predicts when these intuitions can be overcome, either by rewording problems to better align with prior experience or by modifying the learning environment to make alternative strategies more accessible. Together, these results provide a unified account of CRT performance that preserves the insight that CRT errors arise from intuitive reasoning, while situating this intuition within established theories of learning, generalization, and environmental adaptation. It is worth noting that the current approach builds on and extends prior work modeling associative mechanisms that guide reasoning errors in probability and factual judgment tasks such as the Linda problem and the city-size judgment task (Bhatia, 2017). That work has shown how the same associative processes that enable efficient and accurate factual judgment can also produce well-known errors attributed to the misuse of cognitive heuristics.

The success of this approach stems from recent advances in language modeling that allow us to quantify the representations of common stimuli and link them directly to environmental structures and cognitive processes. This is why such approaches are predictive of other types of judgments, including risk, health, and numerical judgments (see Bhatia & Aka, 2022 for a review). This quantification not only increases theoretical precision but also opens new applications, for instance, in educational settings like math tutoring or in applied domains such as consumer finance. The same intuitive missteps that seem amusing or harmless in the context of the CRT can have serious consequences when they occur in the real world. For this reason, researchers must move beyond qualitatively describing these errors in lab environments to developing tools that can predict and prevent them ion the real-world, and this is precisely what a formal modeling approach makes possible.

References

Bago, B., & De Neys, W. (2019). The Smart System 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning, 25*(3), 257–299.

Bhatia, S. (2017). Associative judgment and vector space semantics. *Psychological Review, 124*(1), 1–20.

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

Brañas-Garza, P., Kujal, P., & Lenkei, B. (2019). Cognitive Reflection Test: Whom, how, when. *Journal of Behavioral and Experimental Economics, 82,* Article 101455.

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review, 62*(3), 193–217.

Campitelli, G., & Gerrans, P. (2014). Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. *Memory & Cognition*, *42*(3), 434–447.

Dhami, M. K., Hertwig, R., & Hoffrage, U. (2004). The Role of Representative Design in an Ecological Approach to Cognition. *Psychological Bulletin, 130*(6), 959–988.

Erceg, N., Galic, Z., & Ružojčić, M. (2020). A reflection on cognitive reflection: Testing convergent validity of two versions of the Cognitive Reflection Test. *Judgment and Decision Making, 15*(5), 741–755.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19*(4), 25–42.

Lieder, F., & Griffiths, T. L. (2017). Strategy selection as rational metareasoning. *Psychological Review, 124*(6), 762–794.

Marewski, J. N., & Schooler, L. J. (2011). Cognitive niches: An ecological model of strategy selection. *Psychological Review, 118*(3), 393–437.

Miao, S. Y., Liang, C. C., & Su, K. Y. (2020). A Diverse Corpus for Evaluating and Developing English Math Word Problem Solvers. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 975-984).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *In Proceedings of the 1st International Conference on Learning Representation.*

Patel, N., Baker, S. G., & Scherer, L. D. (2019). Evaluating the cognitive reflection test as a measure of intuition/reflection, numeracy, and insight problem solving, and the implications for understanding real-world judgments and beliefs. *Journal of Experimental Psychology: General, 148*(12), 2129–2153.

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2016). Is the cognitive reflection test a measure of both reflection and intuition? *Behavior Research Methods*, *48*(1), 341–348.

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. *In Proceedings of EMNLP-IJCNLP 2019* (pp. 3982–3992). Association for Computational Linguistics.

Sinayev, A., & Peters, E. (2015). Cognitive reflection vs. calculation in decision making. *Frontiers in Psychology*, *6*, 532.

Szaszi, B., Szollosi, A., Palfi, B., & Aczel, B. (2017). The cognitive reflection test revisited: Exploring the ways individuals solve the test. *Thinking & Reasoning, 23*(3), 207–234.

Welsh, M., Burns, N., & Delfabbro, P. (2013). The cognitive reflection test: How much more than numerical ability? *Proceedings of the Annual Meeting of the Cognitive Science Society, 35*(35).