Modeling the Continued Influence Effect in the Information Environment*

Alexander R. Hough[1], Christopher Fisher[2], Christopher Stevens[1], Taylor Curley[1], Othalia Larue[2], and Christopher Myers[1]

[1] Air Force Research Laboratory, Wright-Patterson AFB, OH 45433, USA
[2] Parallax Advanced Research, Beavercreek, OH 45431, USA

Abstract. Misleading information can have a lasting effect even after it has been corrected or factual discrediting information is learned. This is called the continued influence effect (i.e., CIE) and it can influence judgment at the individual and group level. The CIE has been addressed experimentally, but there are few, if any, cognitive models that specify its mechanisms to make predictions and explain behavior. Here, we discuss relevant literature and modeling efforts, then propose a novel modeling framework for investigating the CIE using cognitive models at the individual level and agent based modeling at the group level. We demonstrate the utility of the modeling framework using simulations which show how the CIE emerges from memory processes and social interactions

Keywords: Continued influence effect · cognitive modeling · ACT-R · instance based learning · agent based modeling · core affect.

1 Introduction

Misinformation and deception were leveraged to successfully mislead adversaries in past conflicts [12, 19]. With advances in technology and rise of social media platforms, the prevalence and spread of misinformation has increased [31]. This creates opportunities to influence individuals and groups by propagating misinformation and exploiting cognitive vulnerabilities. Misinformation can greatly influence decision making and cognition [20], even when the information is later retracted. The lasting effect of data that has been labeled as misleading is called the *continued influence effect*, or *CIE* [14, 18]. Such effects have been identified as having a potential influence on beliefs about topics ranging from climate change [30] to COVID-19 vaccinations [13]. Here, we focus on the CIE, discuss and demonstrate some relevant modeling techniques, and propose future work to model the CIE in a more appropriate context.

* This research was supported by the U. S. Air Force Research Laboratory's 711^{th} Human Performance Wing, Cognition and Modeling Branch. Contents have been reviewed and approved for public release (Distribution A), case number: AFRL-2023-3274 . The views expressed are those of the authors and do not reflect the official guidance or position of the United States Government, the U.S. Department of Defense, the U.S. Air Force, or any of their subsidiaries or employees.

1.1 The Continued Influence Effect

Most theoretical accounts of the CIE focus on the influence of normative episodic memory functioning: Once a piece of information is introduced to long-term memory (LTM), it cannot be erased; the memory can only be re-activated and associated with different information (e.g., a newly-formed memory is false or unreliable) [32]. For information that is later revealed to be false, this can lead to memory errors via competing memory activations [2, 7], recency effects [5], or familiarity-based fluency [6, 29]. Retrieval-failure accounts are also useful in accounting for effects related to the quality of a retraction in the face of misinformation. For example, Lewandowsky and colleagues [18] postulate that corrections often do not replace the coherence of the mental model constructed during the presentation of the original information [32, 14]. In many cases, this provides fewer retrieval pathways to the retraction source compared to the misinformation, leading to higher rates of retrieval failures [26, 21]. Instead, retractions require effective negation "tags" [8] to overcome these memory errors and successfully retrieve the corrected information [15]. However, when there is relational or causal structure between memories (e.g., a narrative), negation is not as effective if does not fit in or provide meaningful structure or causality[14].

The CIE is largely a memory phenomenon that interacts with other cognitive processes, social factors, and affects both individual and group level behavior. To better understand this interaction, make predictions, and explain behavior, we leverage the ACT-R cognitive architecture [1]. ACT-R simulates human-like constraints on performance - such as cognitive bottlenecks - and provides well-validated mechanisms for memory performance and the interaction between cognitive processes.

2 Integrated Modeling Approach

ACT-R is a hybrid cognitive architecture with symbolic and sub-symbolic structures [1]. There are perceptual-motor and memory modules representing systems of the mind. Perceptual-motor modules enable perception of stimuli, actions like pressing buttons, and goal directed behavior. The declarative memory module represents facts as chunks in long term memory and a sub-symbolic component determines their availability. The procedural module represents knowledge about how to do things, represented as condition-action rules. The pattern matcher determines if any rule conditions match the current state and if so, the rule may "fire" and change the state of the model. The behavior of the model is represented as a series of rule firings and state changes. Next, we describe how ACT-R declarative memory and affect mechanisms can capture the CIE at the individual level and introduce an approach to capture the spread of information in groups.

2.1 Individual Level

Instance-based Learning. At the individual level, we use a theory of learning and decision making motivated by ACT-R called instance-based learning

(IBL)[9]. IBL describes how people make decisions in dynamic environments and learn from feedback. According to IBL, an agent encodes each unique experience or *instance* as a chunk (i.e., basic declarative memory unit) consisting of three types of information: 1) Situation is the information available to the agent (e.g., cues), 2) decision is the action taken by the agent, and 3) utility is the outcome of the decision. Utility is usually expressed on an abstract scale, like subjective utility [25], but could be contextualized (e.g. money or points). When making a decision, an agent evaluates each possible action by comparing their *blended values*. A blended value is a noisy representation of an expected utility where outcomes are weighted according to their probability of retrieval. Thus, an important feature of IBL is that decision making is sensitive to memory dynamics such as memory decay, and the frequency and recency of experienced instances. Typically, an agent selects an action corresponding to the highest blended value, but other decision rules could be considered. For instance, if there are no relevant instances, the situation might be judged to be 'atypical' and a heuristic (i.e., procedural memory) may be used to determine a response [10].

In IBL, an instance is a chunk comprised of slot-value pairs containing structured information (i.e., situation, decision, and utility). Each chunk has an activation value representing its ability to be retrieved and is interpreted as the log odds that the information is needed in a given situation [1]. As the activation of a chunk increases, so does the probability and speed of its retrieval. A chunk's activation, A_i, is determined by base level activation, B_i, with some added activation noise, ϵ_i. The base level term is important for IBL as it describes opposing dynamics of learning with experience and forgetting across time. It is stated as:

$$B_i = \log\left(\sum_{j=1}^{n_i} t_{ij}^{-d}\right) \tag{1}$$

where n_i is the number of times chunk i has been used or retrieved, t_{ij} is elapsed time in seconds since the j^{th} retrieval, and $d \in [0, 1]$ is a decay parameter. According to IBL, agents select the option with the highest blended value. A blending mechanism [17] retrieves a value for slot value k by computing a weighted average of all possible values weighted by the their probability of retrieval:

$$\text{bv}_k = \sum_{i \in M} p_i v_{ik}, \tag{2}$$

where M is the index set over all chunks, v_{ik} is the value of slot k in chunk i, p_i is the probability of retrieving chunk i, which is given by the softmax function:

$$p_i = \frac{e^{A_i/\tau}}{\sum_{m \in M} e^{A_m/\tau}}, \tag{3}$$

where τ controls how sensitive the probability weights are to activation. Therefore, if a chunk was created with a high base level activation or was used frequently, it would have more influence in blended values and decision making. This is a natural way to model the CIE within the IBL framework. However, it does not include emotion and its influence on activation.

Core Affect and Weighting. Prior research has demonstrated that emotionally-laden headlines facilitate the spread of misinformation [3]. In its standard form, IBL provides only a partial account of CIE due to its exclusive focus on cognition. Recent research inspired by core affect theory [24], laid the initial groundwork for incorporating affect into cognitive models such as IBL [16]. Core affect focuses on feelings underlying emotion and includes two dimensions: valuation (positive or negative) and arousal (magnitude). A module was developed [16] to compute valuation, V_i and arousal Ar_i, which are both added to the base level term in the activation equation. The current valuation of chunk i at the j^{th} use is based on its previous valuation $V_i(j-1)$ and the difference between the previous valuation and current reward $R_i(j)$ multiplied by the learning rate for valuations av:

$$V_i(j) = V_i(j-1) + \mathrm{av}[R_i(j) - V_i(j-1)]. \tag{4}$$

Arousal is the absolute magnitude of valuation.Valuations are updated each time a chunk is referenced within a time window. Arousal and valuation are included in every activation update and can be used as retrieval cues. The valuation module has three buffers: 1) valuation-filter handles targeted retrieval requests, 2) valuation holds valuation and arousal values for the last retrieved chunk, and 3) core-affect represents the current affective state of the model by calculating and storing the weighted sum of valuations (i.e., core-affect-valuation) and arousal values (i.e., core-affect-arousal) based on all chunks in declarative memory. Valuation affects probability of chunk retrieval via activation, and core affect can be used to reward the model or serve as a proxy for judgments or preferences. For instance, a chunk with associated negative affect could have greater activation and carry more weight in a blended value. This effect could persist over time despite the accumulation of conflicting evidence and impact decision-making. In previous work [16], we computed the Feeling Of Rightness (FOR), which serves as a measure of confidence in current knowledge. It is based on the subjective ease an intuitive answer is generated. The internal reward used in core affect, when knowledge is updated, could influence the FOR through positive reinforcement. A high FOR may lead to superficial or biased answers, and a low FOR may led to questioning knowledge and deeper processing.

Production compilation. By modeling how declarative knowledge is compiled into procedural knowledge rendering memory retrievals no longer necessary, production compilation [1] can explain how original mental models are reinforced and why they are difficult to modify. After compilation, the idea now exists in a form that cannot be verbally retrieved and modified. As a result, retraction or correction is not possible. Instead, a new production with a correction or modification would need to be created to compete with the original.

2.2 Group Level

Information often spreads across a group through direct or social media interactions between individuals. Therefore, it is important to model the social and environmental context in which information is disseminated. At the group level, we use agent based models (ABMs) [4] as a computational framework for modeling the spread of (mis)information within social networks. ABMs are flexible and have been used in a wide variety of fields to model complex systems, including social psychology, sociology, and economics. In ABMs, autonomous *agents*—which could represent cells, individuals, groups, or organizations—are defined by a set of properties (e.g., personality) and relatively simple rules guiding their behaviors. Agents interact with each other and the environment across a series of discrete time steps, causing the state of the system to evolve [28, 4]. Interactions in ABMs are completely bottom-up based on relatively simple rules specified at the agent level. In some ABMs, emergent properties not explicitly encoded in the agents' rules may arise over a series of interactions [28]. One well-known example is flocking behavior of birds [23], which is not programmed into agents, but instead emerges through the interaction of agents following simple rules.

3 Preliminary Modeling Demonstrations

We propose an integrated, multilevel approach to understanding the CIE. IBL theory provides a powerful, well-validated memory theory capable of expressing the memory-based accounts reviewed above. The valuation module provides mechanisms to account for how emotionally charged content affects memory activation and extent of cognitive processing. Group level modeling allows the simulation of this effect over populations of individuals and provides predictions about the spread of CIE through networks of individuals. Integrating these approaches enhances the cognitive fidelity and explanatory power of group-level simulations, allowing agent beliefs and behavior to be guided by memory theory.

3.1 Individual Level Demo

We present a model using IBL without core affect added as an initial demonstration of the approach. The model completes a two alternative forced choice task based on three binary cues [11, 22]. The relationship between the cues and the correct choice was governed by a probabilistic rule such that when a given configuration of cues was present, then the correct answer could be known with a certain probability. This probability was manipulated between subjects (either 80 or 90 percent). In prior studies, we demonstrated this model provided a good fit to empirical data. Here, we wanted to learn the effect of introducing biasing information at the outset of the experiment. Specifically, we introduced instances suggesting the opposite of the correct responses would have strong utility values. We entered two instances in memory that suggested: 1) if the second cue was "2" and a "yes" response was given, it would have a utility of "1", and 2) responding no in this situation would have a utility of "-1". We expected the bias would initially impair the model's performance, but eventually recover as it learned from feedback. In Figure 1, color represents information cost and line type cue validity. The biased model (right column) started with poorer accuracy

Fig. 1. Individual level model results showing bias and partial recovery over time.

(top row), but approached performance of the standard model (left column) by the end. Latency (bottom row) for the biased model was initially faster, as bias led to less consideration of other information in memory. This simple demonstration shows IBL is capable of capturing the phenomenon of the difficulty of correcting an incorrect belief, even after repeated feedback. Follow-on models can be built within this framework to reveal insights about timing of information presentation, in terms of feedback and biasing information.

3.2 Integrated Model Demo

As a proof-of-concept demonstration of CIE using the integrated modeling approach, we developed an ABM simulation using agents with memory systems based on IBL. The goals were to understand whether a misinformation correction intervention could mitigate the CIE and how the timing of the intervention relates to its effectiveness. In our simulation, 100 agents were placed in random cells of a 20×20 grid such that only one agent could occupy a given cell. Each agent was initialized with a chunk containing two slots—one slot, which unbeknownst to the agents, contained either factual information or misinformation, and another slot which contained the source of the information (i.e., agent ID). Each agent had a 10% chance of starting with misinformation and a 90% chance of starting with factual information. At each time step, agents communicated with their neighbors located horizontally, vertically, or diagonally in adjacent cells (i.e., separated by a Chebyshev distance of 1). Communication between agents proceeded in two steps: first, the communicating agent retrieved the most active chunk, and second, the receiving agent encoded a new chunk containing the (mis)information along with its source. After all agents have communicated, we computed a blended value for each agent according to an adaption of Equation 2 for categorical values, and summed the number of agents whose blended value represented misinformation. The resulting value served as a measure of the impact of misinformation within the group of agents. We repeated the steps above 100 times, which constituted a single run of the model.

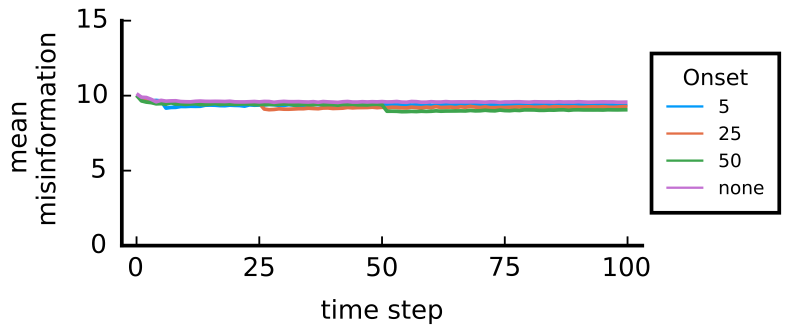

Fig. 2. A demonstration of the CIE at the group-level. Presenting correct information at different time points does not reduce the effect of misinformation.

We devised a simple intervention in an attempt to mitigate the CIE. At one point during the simulation, each agent received corrective information with a $\frac{1}{3}$ probability. If an agent received the corrective information, it was encoded into memory and potentially communicated to neighboring agents. In each condition, the intervention was introduced at the beginning of one of four time steps: 5, 25, 50, or not at all. We assigned each agent the default value for decay of $d = 0.50$ and $\sigma = 0.20$. We set the the retrieval threshold to -10 under the simplifying assumption that retrieval failures are rare. Following the approach used in [9], we used time step (trial) as the unit of time in base-level learning. We ran the model 500 times in each condition and plotted the average values across time steps in Figure 2. The model showed clear evidence of a CIE: misinformation was largely resistant to efforts to correct misinformation. Further inspection of the model revealed that the CIE occured because activation for memories formed early was high and thus attenuated the contribution of subsequent corrections. This result is consistent with prior research showing that corrective information does not overwrite incorrect memories, but instead competes with it [27].

4 Discussion

Developing a comprehensive account of the CIE requires an integrated modeling approach spanning cognition at the individual level and social interactions at the group level. To meet this challenge, we proposed an integrated modeling framework which uses ABM to model social interactions and IBL to develop cognitively plausible agents. One benefit of using IBL at the individual level is that it describes cognitive and emotional processes underlying decision making. ABM complements IBL by providing a framework for modeling social interactions and the spread of (mis)information through social networks.

In line with previous research, our individual level demonstration illustrated that the CIE occurs because corrective information competes with, rather than overwrites, well-established memories. Eliminating the CIE required multiple exposures to corrective information, suggesting effective mitigation may require sustained mitigation over a period of time. Our integrated modeling demonstration illustrated two important points: (1) how the CIE can emerge through a combination of memory mechanisms at the individual level and social interactions at the group level, and (2) how the CIE is largely resistant to interventions designed to present corrective information. Results indicated a small "dose" of corrective information is not sufficient, regardless of timing, suggesting "inoculation" or maintenance doses might be more effective. In the broader social context, mitigating the CIE can be even more challenging. Information exchanged through social interactions tends to be bidirectional and thus mutually self-reinforcing. One implication that could be investigated in future research, is that an effective intervention should target social networks, instead of individuals, and leverage the same information channels and dynamics which facilitated the initial spread of misinformation. The individual demo suggests this intervention should also be multi-dosed and sustained.

4.1 Limitations and Future Work

We plan to refine the proposed integrated modeling approach. For the individual level demonstration, we used an IBL model that was originally developed for a different context. In addition, we discussed core affect mechanisms, but have not yet added to the model. In future work, we plan to: 1) develop a more comprehensive IBL model, 2) include the valuation model to add affective influence on chunk activation and extant of cognitive processing, 3) explore the interplay with declarative and procedural memory, and 4) do some experimental work to better understand cognitive mechanisms underlying the CIE and inform model mechanisms.

The preliminary demonstration for the group level model included basic declarative memory components and each agent only started with one piece of information. The ABM framework provides a solid foundation to model the CIE at the group level and the demonstration suggests it can be scaled in complexity. In future work, we plan to add additional memory components from the individual level model, more pieces of information, and additional dynamics within the group. For instance, agents within the group could have different amounts of information that may or may not be emotionally charged, or have different degrees of influence on other agents to simulate social status or trustworthiness.

References

1. Anderson, J.R.: How can the human mind occur in the physical universe? Oxford University Press, New York, NY (2007).

https://doi.org/10.1093/acprof:oso/9780195324259.001.0001
2. Ayers, M.S., Reder, L.M.: A theoretical review of the misinformation effect: Predictions from an activation-based memory model. Psychonomic Bulletin and Review **5**, 1–21 (1998). https://doi.org/10.3758/BF03209454
3. Baum, J., Abdel Rahman, R.: Emotional news affects social judgments independent of perceived media credibility. Social Cognitive and Affective Neuroscience **16**(3), 280–291 (2021). https://doi.org/10.1093/scan/nsaa164
4. Bonabeau, E.: Agent-based modeling: Methods and techniques for simulating human systems. Proceedings of the national academy of sciences **99**(suppl_3), 7280–7287 (2002). https://doi.org/10.1073/pnas.082080899
5. Ecker, U.K., Lewandowsky, S., Cheung, C.S., Maybery, M.T.: He did it! she did it! no, she did not! multiple causal explanations and the continued influence of misinformation. Journal of memory and language **85**, 101–115 (2015). https://doi.org/10.1016/j.jml.2015.09.002
6. Ecker, U.K., Lewandowsky, S., Swire, B., Chang, D.: Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. Psychonomic Bulletin & Review **18**, 570–578 (2011). https://doi.org/10.3758/s13423-011-0065-1
7. Ecker, U.K., Lewandowsky, S., Tang, D.T.: Explicit warnings reduce but do not eliminate the continued influence of misinformation. Memory & Cognition **38**, 1087–1100 (2010). https://doi.org/10.3758/MC.38.8.1087
8. Gilbert, D.T., Krull, D.S., Malone, P.S.: Unbelieving the unbelievable: Some problems in the rejection of false information. Journal of Personality and Social Psychology **59**(4), 601–613 (1990). https://doi.org/10.1037/0022-3514.59.4.601
9. Gonzalez, C., Dutt, V.: Instance-based learning: integrating sampling and repeated decisions from experience. Psychological Review **118**(4), 523–551 (2011). https://doi.org/10.1037/a0024558
10. Gonzalez, C., Lerch, J.F., Lebiere, C.: Instance-based learning in dynamic decision making. Cognitive Science **27**(4), 591–635 (2003). https://doi.org/10.1207/s15516709cog2704_2
11. Halverson, T., Stevens, C., Fisher, C., Haubert, A., Myers, C.: Balancing confidence and information costs in a diagnostic reasoning task. In: Proceedings of the 16th International Conference on Cognitive Modeling. Madison, WI: University of Wisconsin (2018)
12. Hesketh, R.: Fortitude: The D-Day Deception Campaign. The Overlook Press, Peter Mayer Publishers, Inc (2000)
13. van Huijstee, D., Vermeulen, I., Kerkhof, P., Droog, E.: Continued influence of misinformation in times of covid-19. International Journal of Psychology **57**(1), 136–145 (2022). https://doi.org/10.1002/ijop.12805
14. Johnson, H.M., Seifert, C.M.: Sources of the continued influence effect: When misinformation in memory affects later inferences. Journal of Experimental Psychology: Learning, Memory, and Cognition **20**(6), 1420–1436 (1994). https://doi.org/10.1037/0278-7393.20.6.1420
15. Johnson, H.M., Seifert, C.M.: Updating accounts following a correction of misinformation. Journal of Experimental Psychology: Learning, Memory, and Cognition **24**(6), 1483–1494 (1998). https://doi.org/10.1037/0278-7393.24.6.1483
16. Juvina, I., Larue, O., Hough, A.: Modeling valuation and core affect in a cognitive architecture: The impact of valence and arousal on memory and decision-making. Cognitive Systems Research **48**, 4–24 (2018). https://doi.org/10.1016/j.cogsys.2017.06.002

17. Lebiere, C.: Blending: An ACT-R mechanism for aggregate retrievals. In: Proceedings of the Sixth Annual ACT-R Workshop, George Mason University, Fairfax, VA, USA (1999)
18. Lewandowsky, S., Ecker, U.K., Seifert, C.M., Schwarz, N., Cook, J.: Misinformation and its correction: Continued influence and successful debiasing. Psychological Science in the Public Interest **13**(3), 106–131 (2012). https://doi.org/10.1177/1529100612451018
19. Lewandowsky, S., Stritzke, W.G., Freund, A.M., Oberauer, K., Krueger, J.I.: Misinformation, disinformation, and violent conflict: From iraq and the "war on terror" to future threats to peace. American Psychologist **68**(7), 487–501 (2013). https://doi.org/10.1037/a0034515
20. Loftus, E.F.: Reactions to blatantly contradictory information. Memory & Cognition **7**(5), 368–374 (1979). https://doi.org/10.3758/BF03196941
21. Mayo, R., Schul, Y., Burnstein, E.: "I am not guilty" vs "I am innocent": Successful negation may depend on the schema used for its encoding. Journal of Experimental Social Psychology **40**(4), 433–449 (2004). https://doi.org/10.1016/j.jesp.2003.07.008
22. Myers, C.W., Gluck, K.A., Harris, J., Veksler, V., Mielke, T., Boyd, R.: Evaluating instance-based learning in multi-cue diagnosis. In: NA Taatgen, MK, van Vugt, J. P Borst, & K. Mehlhorn. Proceedings of ICCM 2015. Paper presented at International Conference on Cognitive Science, Groningen, Netherlands (198-199) (2015)
23. Reynolds, C.W.: Flocks, herds and schools: A distributed behavioral model. In: Proceedings of the 14th annual conference on Computer graphics and interactive techniques. pp. 25–34 (1987)
24. Russell, J.A., Barrett, L.F.: Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. Journal of personality and social psychology **76**(5), 805 (1999). https://doi.org/10.1037/0022-3514.76.5.805
25. Savage, L.J.: The foundation of statistics. Wiley (1954)
26. Seifert, C.M.: The continued influence of misinformation in memory: What makes a correction effective? In: Psychology of learning and motivation, vol. 41, pp. 265–292. Elsevier (2002). https://doi.org/10.1016/S0079-7421(02)80009-3
27. Shtulman, A., Valcarcel, J.: Scientific knowledge suppresses but does not supplant earlier intuitions. Cognition **124**(2), 209–215 (2012). https://doi.org/10.1016/j.cognition.2012.04.005
28. Smith, E.R., Conrey, F.R.: Agent-based modeling: A new approach for theory building in social psychology. Personality and social psychology review **11**(1), 87–104 (2007). https://doi.org/10.1177/1088868306294789
29. Swire, B., Ecker, U.K., Lewandowsky, S.: The role of familiarity in correcting inaccurate information. Journal of Experimental Psychology: Learning, Memory, and Cognition **43**(12), 1948–1961 (2017). https://doi.org/10.1037/xlm0000422
30. Treen, K.M.d., Williams, H.T., O'Neill, S.J.: Online misinformation about climate change. Wiley Interdisciplinary Reviews: Climate Change **11**(5), e665 (2020). https://doi.org/10.1002/wcc.665
31. Wang, Y., McKee, M., Torbica, A., Stuckler, D.: Systematic literature review on the spread of health-related misinformation on social media. Social Science & Medicine **240**, 112552 (2019). https://doi.org/10.1016/j.socscimed.2019.112552
32. Wilkes, A., Leatherbarrow, M.: Editing episodic memory following the identification of error. The Quarterly Journal of Experimental Psychology **40**(2), 361–387 (1988). https://doi.org/10.1080/02724988843000168