WG-A: A Framework for Exploring Analogical Generalization and Argumentation

Michael Cooper
Department of Philosophy

Lindsay Fields and **Marc Gabriel Badilla** and **John Licato**
Department of Computer Science and Engineering
Advancing Machine and Human Reasoning (AMHR) Lab
University of South Florida

Abstract

Reasoning about analogical arguments is known to be subject to a variety of cognitive biases, and a lack of clarity about which factors can be considered strengths or weaknesses of an analogical argument. This can make it difficult both to design empirical experiments to study how people reason about analogical arguments, and to develop scalable tutoring tools for teaching how to reason and analyze analogical arguments. To address these concerns, we describe WG-A (Warrant Game — Analogy), a framework for people to analyze analogical arguments based on Bartha's (2010) Articulation Model of analogical argumentation. We carry out two experiments designed to probe WG-A's effectiveness in improving participants' ability to reason about analogical arguments and argumentation in general, and argue that WG-A is a promising approach, though it is in need of further development.

Keywords: analogy; reasoning; generalization; arguments; argumentation; argument analysis; critical thinking

Introduction

Understanding how people reason about and evaluate arguments is a rich area of research, full of competing views on why we reason the way we do, and how to improve it. If there is a consensus among this literature, it is that the ways people tend to evaluate arguments are highly subject to cognitive biases (Walton, 1999; Kahneman, 2011; Mercier & Sperber, 2011; Mercier, 2016; Sperber & Mercier, 2017; Gampa, Wojcik, Motyl, Nosek, & Ditto, 2019)—biases whose effects are prevalent both amongst laypeople and experts such as judges (Guthrie, Rachlinski, & Wistrich, 2001; Chortek, 2013; Wistrich, Rachlinski, & Guthrie, 2015; Rachlinski, Wistrich, & Guthrie, 2015) or medical doctors (Croskerry, 2003b, 2003a; Jenkins & Youngstrom, 2016; Prakash, Bihari, Need, Sprick, & Schuwirth, 2017).

The *myside bias* (Mercier & Sperber, 2011; Mercier, 2016; Sperber & Mercier, 2017) is particularly pernicious because it ostensibly affects our ability to evaluate the quality of arguments with which we are presented, and this effect is also present regardless of cognitive ability (Stanovich & West, 2007). Whether such biases are so prevalent because people are not sure or not willing to restrict their reasoning to that which is relevant to the argument being evaluated, they introduce difficulties both for the empirical study and education of argumentative reasoning.

In this paper, we describe WG-A (Warrant Game - Analogy), a framework and software tool for the evaluation of analogical arguments based on the Articulation Model (AM). AM is a normative model of analogical argumentation that attempts to explain both what a "good" analogy is, and what kinds of dialogical moves can be considered relevant towards assessing an analogical argument (Bartha, 2010). We then report on preliminary studies exploring how the current version of WG-A can be used either as an educational tool or a framework for studying argumentative reasoning, and discuss lessons learned.

Background

An analogical argument consists of propositions divided into source and target domains $\mathbf{S}$ and $\mathbf{T}$. A pair of analogous propositions $(s,t) \in \mathbf{S} \times \mathbf{T}$ is said to be in the *positive analogy* if they have the same truth value, and in the *negative analogy* otherwise. The *hypothetical analogy* is a pair $(h_s, h_t) \in \mathbf{S} \times \mathbf{T}$ such that h_t is the conclusion of the entire analogical argument. As a trivial example: "The sun is round; the moon is round; the sun is very hot; therefore, the moon is very hot." The first two sentences are in the positive analogy, the third is h_s, and the sentence "the moon is very hot" is h_t, the (obviously incorrect) conclusion of the overall analogical argument.

We take as our starting point the *Articulation Model* (AM) of analogical argumentation (Bartha, 2010), whose key idea is that a successful analogical argument explicitly identifies a *prior association* and a *potential for generalization*. The prior association is "a clear connection, in the source domain, between the known similarities [...] and the further similarity that is projected to hold in the target domain" (Ibid.). The prior association has *potential for generalization* when there is a "reason to think that the same kind of connection could obtain in the target domain" (Ibid.). AM describes how these can be made explicit and assessed through a dialogue between an advocate and critic, whose goals are to defend and attack the analogical argument, respectively. In the sun-moon example given above, it is obvious that the connection between being round and being hot is both weak and non-generalizable.

Thus, in a dialogue meant to assess an analogical argument $\mathcal{A}$, a relevant move by a participant is one which contributes to the elaboration or testing of either $\mathcal{A}$'s prior association or potential for generalization. WG-A provides an interface in which participants play the role of critic or advocate, and are only allowed to make moves that have a high probability of being *relevant* (as defined above). We briefly summarize here how WG-A aligns with AM and ensures relevance, but for full discussions, see (Licato & Cooper, 2019, 2020).

A warrant (S. Toulmin, Rieke, & Janik, 1984; S. E. Toulmin, 2003) is a broad principle or rule which shows how an argument's premises permit (or warrant) the inference of its conclusion. They may range from highly formal rules of deductive logic to broad rules of thumb, but they are to be distinguished from premises in that they typically are more generalized (Hitchcock, 2005). WG-A's central assumption is that when given the source and target domains of an analogical argument, the process of elaborating a single warrant which jointly explains the inferences from each domain's facts to its hypothetical is roughly the same task as elaborating a prior association and potential for generalization. For example, consider the analogy in Figure 1. The analogical argument begins with a set of proposition pairs referred to as "facts," each pair containing a proposition from the source domain (left column) and target domain (right column). We will refer to the box labeled 'Facts' on the left side as the *source facts*, and on the right as the *target facts*. The hypothetical analogy is pictured as a pair of propositions in boxes labeled 'Conclusion'. The overall analogical argument is that if the source facts, target facts, and source hypothetical ("cheating on an exam is wrong") are true, then the target hypothetical ("lying on a resume is wrong") follows.

WG-A Gameplay

A WG-A session starts as follows. Two players, filling the roles of advocate and critic, are presented with an interface displaying a pre-selected set of source and target facts and hypotheticals. The advocate is asked to create an initial warrant (referred to as a rule), such that (1) its antecedent is a generalization of the source and target facts, (2) its consequent is a generalization of the source and target hypotheticals, and (3) it serves as a rule which explains both the link from the source facts to source hypothesis, and from the target facts to source hypothesis.

An initial warrant is pictured in the middle column. It consists of an antecedent ("something is dishonest") and consequent ("it is wrong"). Note first that this particular warrant's antecedent does not use all concepts from the available facts; e.g., it doesn't refer to the idea of something which "can be done many ways." The choice of detail to include in the antecedent therefore reflects the warrant's *relevance* to the analogical argument being made (Licato & Cooper, 2019).

Given the structure laid out in Figure 1, a critic can attack links between its parts (labeled **L.1** - **L.5**). An attack on **L.3** requires a counterexample showing that the warrant does not hold (e.g., that white lies are dishonest but not wrong) which might then require the advocate to change the phrasing of the warrant (e.g., changing the warrant's antecedent to "something is dishonest and done for good reasons"), or challenging the attack. The warrant's antecedent must be a generalization of the source and target facts, otherwise the critic can challenge **L.1** or **L.2**; likewise, **L.4** and **L.5** can be attacked if the warrant's consequent does not generalize the hypothetical analogy. Thus, the choice of detail and phrasing used in the warrant is subject to multiple constraints, and the exchange of attacks and edits in response to those attacks leads to iterative improvement of the warrant as the game progresses. A full description of allowed attacks, responses, and other moves is given in Licato and Cooper (2019).

It is important to note that the advocate and critic are not free to directly communicate with each other. They can only perform moves allowed by the rules of the game, and if one player seems to be abusing this in any way, the other has the option to report their conduct to a moderator who will review and respond. This is done to minimize the amount of irrelevance[1] that may occur in an open-ended discussion format.

Experiment 1

Our first experiment was to determine whether using WG-A to evaluate a set of analogical arguments produced any cognitive benefits, as compared to simply engaging in an open-ended dialogue about those same analogical arguments.

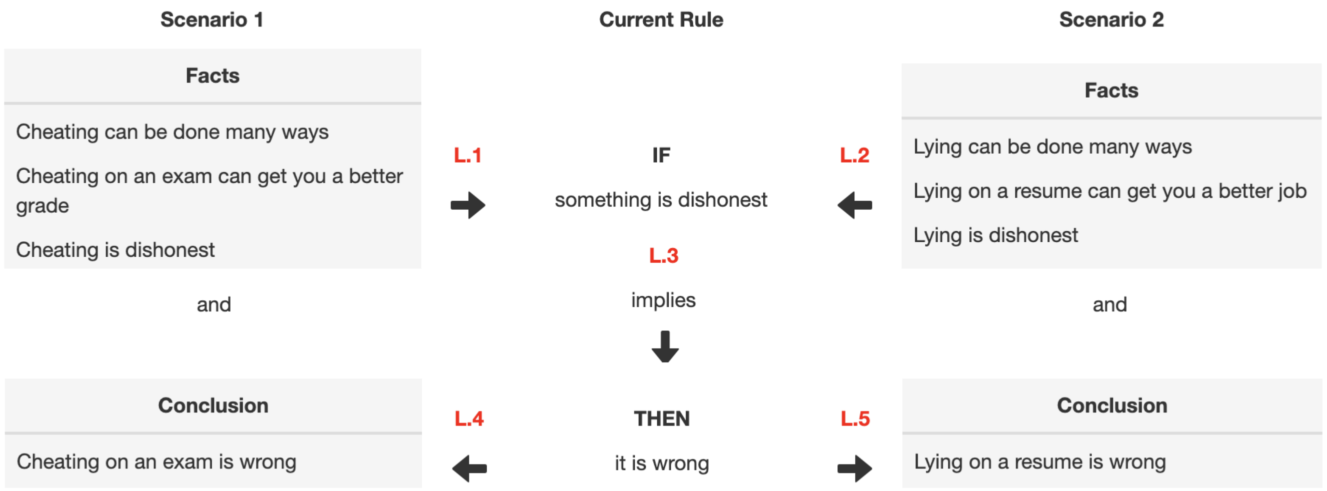
Method

Participants We recruited 64 participants for the first experiment through Amazon Mechanical Turk (MTurk). Each participant was paid a fee of $15 for completion of the study, and offered an additional $2 to complete a follow-up test three days later. One participant took the post-study test two extra times and their extra attempts were excluded from the data. 12 non-participants who initially signed up but did not participate on the day of their main task tried to take the follow-up survey; their results were excluded. 10 participants also completed the main task and the post-study test but did not complete the follow-up survey; these results were excluded, as the majority of statistical analyses used required equal sample sizes, however their inclusion did not impact the results. No demographic data about participants was collected.

Procedure We began our two experiments by asking 750 potential participants on MTurk to fill out their time availabilities, which were used to select a 90-minute session to include both a main task and post-study test. Of the 118 participants selected, 64 logged in and completed their assigned tasks. 44 were assigned to the experimental group and 20 to the control group.

The experimental group began by watching a 15-minute instructional video explaining how to use WG-A. The video contained a password that had to be entered to start the experiment. That completed, they were assigned two WG-A games to work on, with the capability of switching between games while waiting for their turn. In one assigned game they were given the role of advocate, and the other the role of critic. Those who were paired with absent partners were

[1] Following AM and Licato and Cooper (2019), a dialogical move or utterance is considered irrelevant to evaluating an analogical argument if it does not directly lead to a modification of that argument's prior association or potential for generalization, or to the warrant which WG-A claims approximates them.



Scenario 1

Facts

Cheating can be done many ways

Cheating on an exam can get you a better grade

Cheating is dishonest

and

Conclusion

Cheating on an exam is wrong

Current Rule

IF

something is dishonest

implies

THEN

it is wrong

Scenario 2

Facts

Lying can be done many ways

Lying on a resume can get you a better job

Lying is dishonest

and

Conclusion

Lying on a resume is wrong

Figure 1: Screenshot of an example analogical argument + warrant in WG-A

reassigned new games randomly, regardless of their role. The experimental group was guided by the structure of the game to contribute relevant information to the analogy by adding fact pairs, filling in the warrant, and modifying the incomplete analogy that was presented to them.

In order to direct participants towards relevant moves, their interactions were limited to making and justifying a carefully limited set of moves. By limiting the types of moves that participants can make, we hoped to keep them focused and progressing towards complete analogies which list all relevant information. The allowed moves were as follows:

- The advocate first creates the initial rule.
- The advocate can update the rule.
- The advocate or critic can update the facts.
- The advocate or critic can add new fact pairs.
- The critic can attack one of the links.
- The advocate or critic can pass.

Participants were only allowed to pass if 8 moves were already made in their game. Two consecutive passes terminated a game. Participants who completed games before time was up were manually assigned new games to work on. A link to send a report to a moderator was also available, which would suspend the game until a moderator (one of this paper's authors) could review.

Members of the control group were placed into chat rooms in pairs, under the supervision of a moderator. They were presented with information about an incomplete analogy that matched the starting analogical arguments provided to the experimental group. Unlike the experimental group, however, their information was delivered as text, with no graphically-articulated structure and no restrictions on how to improve the analogy. The control group spent their time in unrestricted chatrooms, and given instructions to improve the incomplete analogy that they were given. These instructions were dependent on the scenario but followed this template: "The people who have started this analogy have observed that: *(list of source domain facts)* lead to the conclusion that *(source domain conclusion)*. They have noted that this is similar in some ways to: *(list of target domain facts)* which lead to the conclusion that *(target domain conclusion)*. Please fully explain what the two conclusions have in common, updating their supporting facts as needed to create a good comparison."

Moderators were present in each of the control group's chatrooms. They were not allowed to participate in the discussions in any way, except to (1) provide information about the task and next steps, (2) answer questions about the task itself (not about the topic of discussion), or (3) to remind participants to analyze the argument if both of them were confused or inactive.

Every 15 minutes, the control group was assigned another chat partner and topic randomly, to work on a new analogy. Once participants from either group had been working for 60 minutes, their ongoing games were stopped and they were instructed to spend no more than 15 minutes taking the Test of Scientific Argumentation (TSA) described by Frey, Ellis, Bulgren, Craig-Hare, and Ault (2015). This test measures participants' reasoning abilities with respect to scientific argumentation, in multiple areas including: distinguishing between claims, facts, opinions, and data, or between rebuttals and counter-arguments; determining justification types; identifying qualifier words commonly used in scientific argumentation; and separating scientific from non-scientific claims. For each of these question types, participants were given definitions of the tested words.

Three days after participants' main task and post-study test, they were invited to take the TSA again for an additional $2 fee, if done within 24 hours. In this follow-up test, participants were given the exact same questions as in their post-study test.

Results

With the 44 participants designated to the experimental group, we obtained 18 completed games and 37 which were not completed, but contained significant gameplay. All 20 control group participants contributed to the chat environment. With those 20 participants, we obtained 44 chats with sufficient participation.

The data obtained from Experiment 1 was analyzed to answer the following questions:

- Did participants perform significantly higher in the follow-up test than in the post-study test?
- Did participants in the experimental group perform significantly better than those in the control group?
- Was there a correlation between the level of participation in the game and performance in the test?

To answer the first question, comparisons were performed using a one-tailed, paired Student's t-test. A paired test was deemed appropriate due to the participants being given the same TSA in both the post-study and the follow-up tests. It was found that the experimental group performed significantly higher in the follow-up test than in the initial test, with a p-value of 0.038 and 24 degrees of freedom. This same increase was not observed in the control group, however. On the contrary, the control group performed slightly worse, on average, in the follow-up test, although this decrease was not significant.

To answer the second question, comparisons were performed using a one-tailed, unpaired t-test. It was found that the experimental group performed significantly higher than the control group on the follow-up test, with a p-value of 0.016 and 40 degrees of freedom. However, the groups performed about the same on the initial test. This difference could indicate that the game improves analytical reasoning over time; this theory may be further supported by the control group having no statistical difference in initial versus follow-up test performance. To further test this hypothesis a two-factor ANOVA was calculated, with repeated measures on the test factor as post-study and follow-up test performance is correlated. As shown in Table 1, the results for the experimental group were confirmed to be significantly different from the control group, but the interaction between the group and the test was not significant.

To determine level of participation, we counted the number of game-advancing moves participants in the experimental group performed. For the control group we counted the number of words contributed to the chat. Moves that were not considered game-advancing and, thus, excluded were: passes, acceptances of an opposing player's move, and flags to the moderator. For the control group, we similarly excluded: salutations and regards, asterisked corrections of typos, explanations of technical issues, messages directed to the moderator, and emojis. With these stipulations, we found no significant correlation between participation and test performance for either group, using a Pearson correlation coefficient.

Experiment 2

Our second experiment also divided participants into an experimental group which used WG-A and a control group which used open dialogue. However, whereas Experiment 1 used the TSA to determine whether participants' *general* scientific argument analysis ability was affected, Experiment 2 used a different set of questions designed to assess their ability to analyze a single *analogical* argument.

Method

Participants As with Experiment 1, participants were recruited from MTurk and asked for their availabilities. 89 participants were scheduled for a 90-minute session, of which 48 logged in. Each participant was offered $15 to participate in the study. No demographic data on participants was collected.

Procedure Of those who logged in to the scheduled session, 34 participants were assigned to the experimental group and 14 to the control group. All procedures for the experimental and control groups then were the same as in Experiment 1, save for the post-study test they took.

The Experiment 2 post-study test presented participants with one of six deliberately flawed analogical arguments, as in Figure 2. They were asked to list strengths and weaknesses of the argument, in bullet-point form so that they could be counted. They were also asked to numerically rate the validity of the argument on a five-point scale, and to express their confidence in this rating on a three-point scale (Figure 3). Finally, they were asked to think back about the analogies they examined during the main task, and instructed to evaluate their satisfaction with: the quality of the analysis they performed, the relevance of the things said while doing that task, and whether their understanding of the topics involved changed. Participants were told that the top performers on the Experiment 2 post-study test would be rewarded with an additional bonus payment (Figure 3).

Results

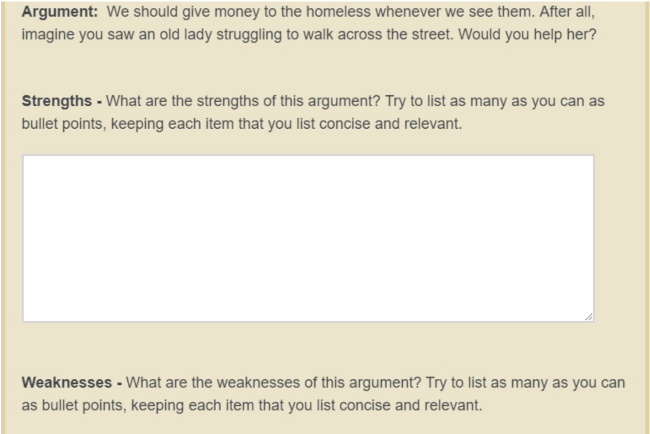
With the 34 participants designated to the experimental group, we obtained 22 completed games and a further 22 which were not completed, but contained significant gameplay. With the 14 control group participants, we obtained 30 chats with sufficient participation.

The data obtained from Experiment 2 was analyzed to answer the following questions:

- Did participants in the experimental group give significantly different responses to analytical questions than those in the control group?
- Was level of participation in the game correlated with analytical response?

Table 1: Experiment 1 ANOVA.

Source	SS	df	MS	F	P
Between subjects	1334.57	41			
Group	88.95	1	88.95	2.86	0.098586
Subjects within Group	1245.62	40	31.14		
Within subjects	199	42			
Test	5.76	1	5.76	1.26	0.268344
Group × Test	10.98	1	10.98	2.41	0.128440
Test × Subjects within Group	182.26	40	4.56		
Total	1533.57	83			

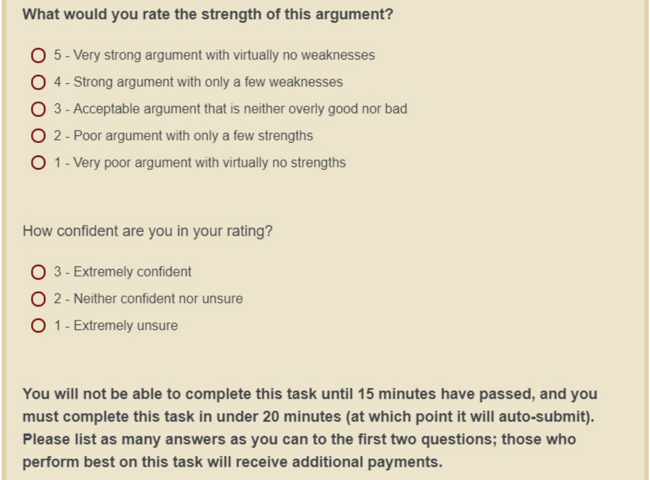


Argument: We should give money to the homeless whenever we see them. After all, imagine you saw an old lady struggling to walk across the street. Would you help her?

Strengths - What are the strengths of this argument? Try to list as many as you can as bullet points, keeping each item that you list concise and relevant.

Weaknesses - What are the weaknesses of this argument? Try to list as many as you can as bullet points, keeping each item that you list concise and relevant.

Figure 2: Participants were asked to list strengths and weaknesses of a given analogical argument.



What would you rate the strength of this argument?

○ 5 - Very strong argument with virtually no weaknesses
○ 4 - Strong argument with only a few weaknesses
○ 3 - Acceptable argument that is neither overly good nor bad
○ 2 - Poor argument with only a few strengths
○ 1 - Very poor argument with virtually no strengths

How confident are you in your rating?

○ 3 - Extremely confident
○ 2 - Neither confident nor unsure
○ 1 - Extremely unsure

You will not be able to complete this task until 15 minutes have passed, and you must complete this task in under 20 minutes (at which point it will auto-submit). Please list as many answers as you can to the first two questions; those who perform best on this task will receive additional payments.

Figure 3: Participants were asked to rate the overall strength of a given analogical argument.

- Was there correlation between the number of perceived strengths and weaknesses in an argument and the argument's overall perceived strength?

To answer the first question, comparisons were performed using a two-tailed Student's t-test. No significant differences were found between the control and the experimental groups' values for any of the quantitatively rated responses.

As in Experiment 1, to determine level of participation we counted the number of game-advancing moves or the number of words contributed to the chat environment. No significant correlation was found between level game participation and analytical responses for either group, using a Pearson correlation coefficient. However, a positive correlation was found between both groups' self-reported understanding of the task and their rating of argument strength. In the case of the control group, this correlation was particularly strong ($p = 0.003$). This could be an indication of participants' confirmation bias; e.g., participants may have believed if an argument appeared strong, they must have understood the task.

Since the weaknesses and strengths of each argument were provided in bullet-point format, we were able to count these, and compare them with the numerical rating of overall argument strength. Using a Pearson correlation coefficient, a strong negative correlation was found between arguments' weakness counts and their strength ratings. ($p = 0.00003$). However, a contrasting correlation did not hold, in general, between strength counts and strength ratings. When considering these same factors, separated by group, it was found that the correlation held roughly equally in both directions for the control group, but a correlation between strength counts and strength ratings was completely nonexistent for the experimental group. Paired with the strong negative correlation between weaknesses counts and strength ratings, this may indicate that participants in the experimental group were more inclined to evaluate arguments by focusing on their weaknesses (i.e. counterarguments against them). Further experimentation would be needed to confirm this.

Additionally, we observed a negative correlation between the ratio of weakness counts to strength counts, and strength ratings, with a p-value of $3x10^{-7}$. In direct contrast to our observations for strength counts, the experimental group was statistically likely to say that an argument with a high ratio of strength count to weakness count was strong (with a p-value of 0.0001), as well as that one with a high ratio of weakness count to strength count was weak (with a p-value of $6x10^{-7}$). Conversely, the control group had a much weaker correlation in both regards, with a p-value of 0.047 for the ratio of strength count to weakness count and 0.002 for ratio of weakness count to strength count. This may indicate that the experimental group displayed an ability to distinguish between an argument having several strong aspects and that same argument being more strong than weak. The ability to distinguish arguments in this manner was not observed in the control group.

General Discussion and Future Work

This paper described the first attempt to empirically study WG-A, focusing on whether its use has any short- or long-term effects on participants' abilities to reason about analogical arguments and argumentation in general. Our results suggest it has potential in at least two areas: (1) to study and teach analogical inference, generalization, and argumentation, and (2) as a framework for the development of automated reasoning. However, it is clear that much more work is needed in all three of these areas, which we now discuss.

The results of Experiment 1 suggest the use of WG-A, when compared with open-ended discussion, improved performance on the TSA. But it is not known why this effect seemed delayed: why was the performance difference between experimental and control groups significant in the follow-up test taken three days after the task, but not in the post-study test taken immediately after? It may be that using WG-A increases participants' interest level in the practice of finding supporting arguments and counterarguments; however, this does not appear to have been reflected in the test results from Experiment 2. Future work should explore whether the effect reported in this paper is robust across other measures of argumentative reasoning.

Licato and Cooper (2019) suggest that WG-A's minimal need for moderation and restricting of allowed communications between players makes it ideal for training artificially intelligent cognitive systems, e.g. to play WG-A against other algorithms, to play against other people, or to use internally as a normative model of analogical argumentation. Indeed, of the 78 participants who played WG-A games, there were only 7 reports to the moderators, only one of which was about an opponent. Furthermore, this was the only case in which a player was clearly acting in bad faith: the player's opponent reported that they were passing and not contributing, instead using the additional detail prompts to urge their opponent to pass so that the game would be over quickly. This suggests that WG-A is scalable, as a large number of games can be played with minimal need for human supervision.

The antecedent and consequent of the warrant are meant to be generalizations of the source and target domains' facts and conclusions, respectively. But future work can explore whether the additional constraints that WG-A places on these generalizations still can be captured by existing cognitive models of analogical generalization (Kuehne, Forbus, & Gentner, 2000; Forbus, Klenk, & Hinrichs, 2009; Hummel, 2001; Doumas, Hummel, & Sandhofer, 2008).

The participants in the experimental group seemed to have particular difficulty understanding that the warrant should be a generalization of the fact pairs in the scenario. More can be done on the presentation side to ensure that warrants are seen as generalizations of their respective components. For example, highlighting the words that are shared between the relevant section of the warrant and the facts or conclusion may help participants to better understand their relationships to one another. In addition, we discussed offering suggestions for new warrant edits drawn from the fact pairs. By suggesting ways to draw from the fact pairs and conclusion, the participants might better understand the types of considerations they need to make in making warrant edits.

Acknowledgments

This material is based upon work supported by the Air Force Office of Scientific Research under award numbers FA9550-17-1-0191 and FA9550-18-1-0052. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the United States Air Force.

References

Bartha, P. F. (2010). *By Parallel Reasoning: The Construction and Evaluation of Analogical Arguments*. Oxford University Press.

Chortek, M. (2013). The psychology of unknowing: Inadmissible evidence in jury and bench trials. *The Review of Litigation*, *32*(117).

Croskerry, P. (2003a, 2018/04/15). Cognitive forcing strategies in clinical decisionmaking. *Annals of Emergency Medicine*, *41*(1), 110–120. Retrieved from `http://dx.doi.org/10.1067/mem.2003.22` doi: 10.1067/mem.2003.22

Croskerry, P. (2003b). The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine*, *78*(8).

Doumas, L. A., Hummel, J. E., & Sandhofer, C. (2008). A Theory of the Discovery and Predication of Relational Concepts. *Psychological Review*, *115*(1), 1-43.

Forbus, K., Klenk, M., & Hinrichs, T. (2009). Compaion Cognitive Systems: Design Goals and Lessons Learned So Far. *IEEE Intelligent Systems*, *24*(4), 36-46.

Frey, B., Ellis, J., Bulgren, J., Craig-Hare, J., & Ault, M. (2015, 01). Development of a test of scientific argumentation. *Electronic Journal of Science Education*, *19*.

Gampa, A., Wojcik, S. P., Motyl, M., Nosek, B. A., & Ditto, P. H. (2019). (ideo)logical reasoning: Ideology impairs sound reasoning. *Social Psychological and Personality Science*. Retrieved from https://doi.org/10.1177/1948550619829059 doi: 10.1177/1948550619829059

Guthrie, C., Rachlinski, J. J., & Wistrich, A. J. (2001). Inside the judicial mind. *Cornell Law Review*, *86*(4).

Hitchcock, D. (2005). Good reasoning on the toulmin model. *Argumentation*, *19*(3), 373-391.

Hummel, J. E. (2001). Complementary Solutions to the Binding Problem in Vision: Implications for Shape Perception and Object Recognition. *Visual Cognition*, *8*(3), 489-517.

Jenkins, M. M., & Youngstrom, E. A. (2016). A randomized controlled trial of cognitive debiasing improves assessment and treatment selection for pediatric bipolar disorder. *Journal of Consulting and Clinical Psychology*, *84*(4), 323-333.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Girous.

Kuehne, S. E., Forbus, K. D., & Gentner, D. (2000). Seql: Category learning as progressive abstraction using structure mapping. In *Proceedings of cogsci.*

Licato, J., & Cooper, M. (2019). Evaluating relevance in analogical arguments through warrant-based reasoning. In *Proceedings of the european conference on argumentation (eca 2019).*

Licato, J., & Cooper, M. (2020). Assessing evidence relevance by disallowing direct assessment. In *Proceedings of the 12th conference of the ontario society for the study of argumentation.*

Mercier, H. (2016). The argumentative theory: Predictions and empirical evidence. *Behavioral and Brain Sciences*, *20*(9), 689-700.

Mercier, H., & Sperber, D. (2011). Why do humans reason? arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*(2), 57-74.

Prakash, S., Bihari, S., Need, P., Sprick, C., & Schuwirth, L. (2017, Feb 08). Immersive high fidelity simulation of critically ill patients to study cognitive errors: a pilot study. *BMC Medical Education*, *17*(1), 36. Retrieved from https://doi.org/10.1186/s12909-017-0871-x doi: 10.1186/s12909-017-0871-x

Rachlinski, J. J., Wistrich, A. J., & Guthrie, C. (2015). Can judges make reliable numeric judgments? distorted damages and skewed sentences. *Indiana Law Journal*, *90*.

Sperber, D., & Mercier, H. (2017). *The enigma of reason* (Audible Audio Edition ed.). Tantor Audio.

Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking & Reasoning*, *13*(3), 225-247. Retrieved from https://doi.org/10.1080/13546780600780796 doi: 10.1080/13546780600780796

Toulmin, S., Rieke, R., & Janik, A. (1984). *An introduction to reasoning* (2nd ed.). New York, New York: Macmillan Publishing Company.

Toulmin, S. E. (2003). *The uses of argument (updated edition)* (Updated ed.). Cambridge University Press.

Walton, D. (1999). *One-sided arguments: A dialectical analysis of bias*. State University of New York Press.

Wistrich, A. J., Rachlinski, J. J., & Guthrie, C. (2015). Heart versus head: Do judges follow the law or follow their feelings?. *Texas Law Review*, *93*(4), 855 - 923.