Measuring Quality of General Reasoning

Alexandru Marcoci[1,*], Margaret E. Webb[2], Luke Rowe[3], Ashley Barnett[4], Tamar Primoratz[4], Ariel Kruger[4], Benjamin Stone[2], Michael L. Diamond[2], Morgan Saletta[4], Tim van Gelder[4], Simon Dennis[2]

[1] Centre for Argument Technology, University of Dundee, AMarcoci001@dundee.ac.uk
[2] Melbourne School of Psychological Sciences, University of Melbourne
[3] School of Education, Australian Catholic University
[4] Hunt Laboratory for Intelligence Research, University of Melbourne

Abstract

Machine learning models that automatically assess reasoning quality are trained on human-annotated written products. These "gold-standard" corpora are typically created by prompting annotators to choose, using a forced choice design, which of two products presented side by side is the most convincing, contains the strongest evidence or would be adopted by more people. Despite the increase in popularity of using a forced choice design for assessing quality of reasoning (QoR), no study to date has established the validity and reliability of such a method. In two studies, we simultaneously presented two products of reasoning to participants and asked them to identify which product was 'better justified' through a forced choice design. We investigated the criterion validity and inter-rater reliability of the forced choice protocol by assessing the relationship between QoR, measured using the forced choice protocol, and accuracy in objectively answerable problems using naive raters sampled from MTurk (Study 1) and experts (Study 2), respectively. In both studies products that were closer to the correct answer and products generated by larger teams were consistently preferred. Experts were substantially better at picking the reasoning products that corresponded to accurate answers. Perhaps the most surprising finding was just how rapidly raters made judgements regarding reasoning: On average, both novices and experts made reliable decisions in under 15 seconds. We conclude that forced choice is a valid and reliable method of assessing QoR.

Keywords: Reasoning, quality of reasoning, forced choice.

Introduction

Forced choice is a standard experimental paradigm common to psychophysics and cognitive psychology, particularly in the assessment of processes such as perception and memory (Link, 1975; Ratcliff & Smith, 2004; Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006), as well as in machine learning (Cao et al., 2007). Typically, a forced choice protocol involves presenting two or more test items simultaneously, with a decision required of the participant (e.g., which item is brighter). Recently, this experimental paradigm has been extended to measuring argument quality. Multiple machine learning models for automatically judging the quality of reasoning (QoR) in written products have been proposed which are trained on short, labelled arguments. To generate these gold-standard corpora, human annotators were instructed to choose between argument pairs (usually on a similar topic and stance), assessing convincingness (Habernal & Gurevych, 2018), evidence strength (Gleize et al., 2019), or which would be adopted by more people (Toledo et al., 2019; Gretz et al., 2020).

The aim of the current research was to explore the criterion validity of the forced choice procedure with minimal instructions in assessing QoR. Criterion validity assesses whether a measure is positively related to other measures one would expect it to be related to. The SWARM project team[1] (which included all authors on this paper) constructed a corpus of 279 arguments (Avg=162 words, SD=132 words) in support of answers to a wide range of reasoning problems with normatively correct solutions. These ranged from standard GMAT/LSAT questions to novel geolocation tasks familiar to human rights researchers attempting to confirm the authenticity of video or photographic footage (see Table 1). Problems were selected according to an established group task taxonomy that originated from social psychology (see McGrath 1984). They involved differing degrees of abstract reasoning (e.g., Raven's matrices), while others relied on general (e.g., integrative reasoning) or domain-specific knowledge with engineering and technology (e.g., Bayesian reasoning). Therefore, we investigated the applicability of the forced choice design to the measurement of QoR across a broad range of domains of reasoning.

Within this problem-selection paradigm, we first expected normatively correct answers would be accompanied by better reasoned rationales in support of their answers. Second, we expected products generated by larger teams to produce answers that were more accurate and better reasoned. Finally, we expected the correlations between objective accuracy and quality of reasoning to be stronger for expert than for novice raters.

In two experiments, we instructed participants to choose what they perceived to be the better-reasoned rationale out of pairs of written arguments supporting different answers to the same problem. Study 1 used an MTurk sample, and Study 2 used an expert sample, composed of people with appropriate training in judging reasoning (see below).

Study 1: Assessing Forced Choice using Novice Raters

In Study 1, we measured criterion validity by assessing if accuracy and team size affected whether a rationale was selected as better reasoned through a forced choice design. We

[1] https://www.imperial.ac.uk/security-institute/research/data-processing-and-algorithms/swarm/

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society*.

pre-registered our hypotheses on the Open Science Registry (see https://osf.io/re5ha). We used the pre-registration template provided by AsPredicted.org and made available at the Open Science Framework (see https://osf.io/m3spx/). We hypothesized that: (1) products resulting in more accurate solutions will be associated with rationales that are chosen more often in forced choice comparisons; and (2) teams with larger numbers of individuals will produce better justified products compared to teams with smaller numbers.[2]

Participants

MTurk raters (N=218) completed the Human Intelligence Tasks (HITS) at the rate of USD 10/hr. Each pair was evaluated by exactly 3 raters.

Materials

Rationales were produced by teams in IARPA's Crowdsourcing Evidence, Argumentation, Thinking and Evaluation (CREATE) program. An email invitation was sent to 4179 members of a pool managed by the Smartly-assembled Wiki-style Argument Marshalling (SWARM) project (van Gelder et al., 2020), of which N=233 consented to participate. They were assigned to teams of varying sizes in two production protocols, as follows: 4 teams of 5 people, 6 teams of 10 people, 4 teams of 15 people, and 4 teams of 21 people, split evenly across protocols. Participants were given 48 hours (February - March 2019) to solve 19 problems (Table 1). Two problems, however, were later removed from the final dataset as they were mistakenly presented to groups twice (e.g., Logical Reasoning 1 and 2 were the same, Raven's Matrices 1 and 2 were the same). The final dataset of problems was therefore based on 17 unique items. These were selected to afford different types of collective reasoning in group performance contexts. Our item-sampling procedure was guided by studies that had previously attempted to measure collective intelligence in human groups (see Engel et al., 2014; Riedl et al., 2021; Woolley et al., 2010). In these studies, a group-IQ test battery was sampled according to McGrath's task circumplex, a group-task taxonomy that divides group tasks into four qualitatively distinct quadrants with eight subdimensions: generate (creative, planning), choose (intellective, decision-making), negotiate (cognitive conflict, mixed motives), and execute (performances / actions, contests / competitions) (see McGrath, 1984, p. 61). Each team submitted a single answer to each problem, though not all teams completed all tasks (and some answers were excluded due to poor quality). In total, 279 rationales of between 3-856 words were collected.

Table 1. Problems used in Study 1 and Study 2

Problem	**Description**	**#products**
Verbal Comprehension 1 (VBC_1)	Tests comprehension of written text (GMAT, 2018)	18
Verbal Comprehension 2 (VBC_2)		14
Geolocation 1 (GEO_1)	Asks for the location and time of a given photo (in-house)	16
Geolocation 2 (GEO_2)		12
Geolocation 3 (GEO_3)		14
Critical Reasoning 1 (CR_1)	Tests ability to critique an argument (GMAT, 2018)	17
Critical Reasoning 2 (CR_2)		13
Object Identification (OID_1)	Participants are required to identify an object (in-house)	16
Integrative Reasoning (IR_1)	Tests ability to draw the correct conclusions from data (Manhattan Review, 2012)	17
Document Identification (DocID_1)	Participants must correctly identify the source of the text (in-house)	15
Syllogisms Problem (Syl_1)	Tests ability to identify consequences of deductive syllogisms (Ennis et al., 1985)	17
White-team Checkers (Che_1)	Based on 5 preceding checkers moves, participants need to correctly predict the 6th move based on a real game (in-house)	14

[2] We pre-registered a third hypothesis regarding the difference in accuracy between production protocols. This does not have any bearing on assessing the criterion validity of a quality of reasoning measure and we don't report on it in this manuscript.

Logical Reasoning 1 (LR_1)	Tests understanding of logical principles (LSAT, 2015)	17
Logical Reasoning 2 (LR_2)		9
Raven's Matrices 1 (Mx_1)	A validated test of fluid intelligence and spatial reasoning (Raven, 1998)	16
Raven's Matrices 2 (Mx_2)		13
Simple Probabilistic (Bayesian) Reasoning (Bay_1)	Tests capacity to correctly update probabilities based on evidence (Mandel, 2015)	15
Complex Probabilistic (Bayesian) Reasoning (Bay_2)	Tests ability to extract relevant probabilistic information and use it in a Bayes net to update probabilities (Lagnado, Liefgreen & Pilditch, 2017)	13
Estimation Problem (Est_1)	To answer correctly the team must correctly estimate the number of candies in the jar (in-house)	13

Procedure

Raters were provided with the following instructions:

A set of complex questions were presented to teams of individuals to solve within 48 hours. Teams were asked to both: 1) Provide the correct answer to each problem, and 2) To provide the background rationale for their answer. In the current HIT, we will 1) Present you with the problems participants were shown, and 2) Ask you to evaluate the reasoning of the answers teams generated. Two pieces of rationale will be presented at the same time: Your task is to decide which team you think justified their answer best by clicking on your preferred rationale.

Raters were then presented with a randomly allocated problem statement (see supplementary materials on OSF). Once they read through the problem statement, raters were presented with two randomly selected rationales corresponding to the problem statement. The rationale that was deemed to be "better justified" was then chosen by the rater. Once the choice was made, they were presented with two more randomly drawn rationales. On average each rater saw 26.4 pairs of rationales (SD = 31.1). This amounted to a total of 1,915 comparisons and choices. Raters were not informed how accurate it was and in many cases the responses were equally accurate. Data collection took place in May 2019.

Results

Accuracy To assess the relationship between accuracy and the forced choice measure of quality we counted the number of times the team whose answer was closest to the correct solution produced the rationale that the majority of raters chose (Figure 1). For instance, for the GEO problems, teams were required to produce a set of GPS coordinates. The team whose coordinates were closest to the correct answer were deemed to be most accurate. A match was recorded if the majority of the raters chose their rationale. Note that answers that were equally correct were not considered for this analysis (e.g., when two teams provided the correct answer in a multiple-choice question). Participants chose the rationale supporting the more accurate solution 68% of the time (SD = 1%).

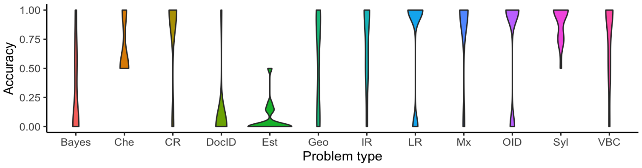

Figure 1. Accuracy by problem-type

Comparison Between Team Sizes Larger teams produced rationales that were more likely to be chosen compared to teams with fewer members (Table 2). For example, when comparing products created by teams with 21 allocated members to products created by teams with 5 allocated members, participants chose products created by teams with 21 members 82% (SD = .02) of the time, which equates to an effect size of 1.29.

Inter-rater Reliability The percent agreement between raters was 70.58% (95% CI = 1.18). Chance agreement is 50%, so performance is significantly and substantially better than chance, although far from perfect.

Response Time (Exploratory) While raters must read products upon first presentation, the majority of comparisons were between pairs of products that raters had read previously and judgements were made quite rapidly. The median reaction time (RT) per comparison was just ~9 seconds (mean response time = 29.9 seconds; SD = 100.07). The median response times per problem are outlined in Table 3. These results indicate that forced choice assessments of QoR can be completed within seconds. The correlation between median RT and the probability of choosing the rationale that corresponded to the more accurate answer is -.11; that is, we observed a small tendency for quicker judgments to be more accurate.

Discussion

Determining QoR is inherently subjective and context dependent (Woods, 2013). Even when provided with detailed guidance, human raters tend to exhibit judgements that have low reliability (Wachsmuth et al., 2017). Study 1 establishes that a forced choice design can be used to evaluate QoR. Prompting novice raters to make assessment QoR in relation to similar products tends to facilitate valid, reliable, and efficient judgments that align with various dimensions of accuracy. This finding confirms our pre-registered hypotheses that more accurate solutions tend to be associated with the chosen rationale in a forced choice comparison.

A written rationale with more accurate reasoning was significantly more likely to be chosen over one with less accurate reasoning, and this trend was relatively strong even among individual raters with no prior training and only minimal guidance. Furthermore, these trends were observed across a wide range of problems with different kinds of reasoning and different levels of difficulty. Indeed, while only 7% of the answers to the Bay_1 problem were correct, raters nevertheless selected the more accurate Bay_1 rationale in 55% of cases. For Doc_ID we only found 17% correct answers, but raters achieved 77% accuracy.

Second, we expected that many of the problems would require substantial outside knowledge and would follow a "truth wins" schema. This schema can be exemplified in the context of the geolocation questions (GEO_1, 2, and 3). To identify an image as being taken in Sao Paolo required familiarity with urban architecture in Brazil and even the distinct aspect of phone booths and streetlights in Sao Paolo. Nevertheless, once somebody correctly identifies the location of the image, they need only to send a Google Maps Street View link to their team members to convince them. Therefore, we reasoned that the probability of a given group member knowing or discovering the solution would be greater in larger compared to smaller teams, which in turn would suggest that larger teams would outperform smaller ones both in accuracy and in the quality of their rationales. This was reflected in our second hypothesis, which was supported by the results: both novices and experts consistently selected the reports generated by larger teams as being better reasoned, amounting to substantial effects.

Finally, our secondary analysis found raters can make relatively accurate forced choice comparisons in a brief amount of time. For example, the median reaction time was ~9 seconds for MTurk participants; although, it should be noted that this trend is not obvious when using the statistical mean because the distribution was highly skewed by the initial reading of the products, which typically takes most participants significantly longer than 9 seconds

Study 2: Assessing Forced Choice using Expert Raters

In Study 2, we investigated the performance of expert raters with no training and no calibration.

Participants

"Expert" raters (N = 6) were selected on the following criteria: 1) completed or currently completing a postgraduate degree in logic or the psychology of reasoning, and 2) have teaching experience (and had graded assessments) in logic. We recruited 5 postdoctoral fellows and 1 advanced PhD student. On average, the experts had 4.83 peer-reviewed articles (SD=5.27) and taught 13.16 undergraduate courses (SD=7), 4.66 of which in logic (SD=4.36). Raters were compensated at approximately AUD 40/hr. Each pair was evaluated by 2 raters.

Methods

The materials, procedure, and measures were as in Study 1, with one exception. For this study, we selected only 9 problems (149 products) and constructed exhaustive comparisons (1,162 unique comparisons).

Results

Accuracy Experts chose the rationale closer to the normatively correct solution 78% of the time (SD=2%), which corresponds to an effect size of 0.82 (SD=.17). See Table 3 for further details.

Comparison Between Team Sizes As in Study 1, we found that larger teams produced rationales that were more likely to be chosen (Table 2); that is, experts were more likely to select using the forced choice methodology the rationales that were generated by the larger teams.

Table 2. Bayesian probability estimates of choosing products created by the team with higher numbers of allocated members, by MTurk and Expert raters. Below the diagonal line are mean probabilities (and SD); above the diagonal line are effect sizes (and SD). Responses by MTurk raters and Expert raters are the top and bottom halves of the table, respectively.

MTurk

	21	15	10	5
21	-	.40 (.07)	.54 (.15)	1.29 (0.22)
15	.61 (.01)	-	.40 (.07)	.87 (.17)
10	.65 (.02)	.61 (.01)	-	.74 (.16)
5	.82 (.02)	.73 (.02)	.70 (.02)	-

Experts

	21	15	10	5
21	-	.43 (.22)	.54 (.15)	1.47 (.24)
15	.62 (.03)	-	0.07 (.14)	.90 (.26)
10	.65 (.02)	.52 (.02)	-	.78 (.17)
5	.85 (.02)	.74 (.03)	.71 (.02)	-

Inter-rater Reliability The percent agreement between the raters was 80.98% (95% CI = 2.26). As with the novice raters, the reliability is significantly above chance although not excellent. The percent agreement is significantly better for experts than for novices (70.58%), as one would expect - adding to the case for the criterion validity of the procedure. However, the difference is perhaps not as substantial as one might have expected. We will return to this point in the discussion.

Response Time (Exploratory) The response time for comparisons was slightly longer for experts compared to MTurk raters (grand median response times were ~14.4 vs. 9 seconds, respectively); however, experts still made their comparisons very quickly (mean response time = 26.77, SD = 40.69). Comparisons of median response time broken down by problem types are presented in Table 3.

Table 3. Descriptive statistics by problem for average proportion correct (%Corr), median response time in seconds (RT), and probability that forced choice responses would reflect proximity to the correct answer for MTurk and Expert raters (Acc)

	MTurk			Expert	
Problem	**%Corr**	**RT**	**Acc**	**RT**	**Acc**
Avg	**0.61**	**8.97**	**0.68**	**14.39**	**0.78**
Bay_1	0.07	10	0.55	16.25	0.45
Cbay_1	0.51	9.5	0.59	-	-
Che_1	0.67	9	0.49	-	-
CR_1	0.78	8	0.95	18.5	0.93
CR_2	0.86	6	0.36	-	-
DocID_1	0.17	8	0.77	-	-
Est_1	0.23	10	0.51	-	-
GEO_1	0.62	9.5	0.65	14	0.63
GEO_2	0.43	10	0.69	-	-
GEO_3	0.57	10	0.8	-	-
IR_1	0.74	7	0.78	14	0.79
LR_1	0.61	10	0.64	9.5	0.70
Mx_1	0.94	6.5	1	10.5	0.93
OID_1	0.75	7	0.64	11	0.85
Syl_1	0.9	12	0.83	24.75	0.88
VBC_1	0.75	9	0.75	11	0.84

VBC_2	0.79	11	0.5	-	-

Discussion and Conclusions

Experts chose the product that supported the more accurate answer substantially more often than novices (.78 as compared to .68). One notable exception was the Bayes network problem (Bay_1) in which accurate solutions were scarce (e.g., 7%) compared to other problem-types. This not only reduced the number of accurate written products among which QoR could be assessed, but may also have undermined the expert raters' capacity to clearly discriminate better- from worse-reasoned rationales. By removing the Bay_1 as an outlier, expert accuracy would increase from .78 to .82.

Experts were also slower to respond (mean of median reaction times was ~14.4 for experts versus ~9 seconds for MTurkers). As mentioned above, however, the median reaction time and the probability that the more accurate product was chosen were negatively correlated for the MTurk participants, suggesting that speed alone was not the reason why they were less likely to choose the most accurate answer. By restricting our focus to just those problems that were presented to experts, this correlation shifts from -0.11 to -0.35. By contrast, this correlation was 0.07 for the experts.

While experts achieved higher percent agreement than novices and, in both cases, performance was significantly above chance, the agreement was not particularly strong. Agreement depends on the consistency with which the raters are addressing the same construct, and on the discriminability of the choices. It may be that many of our products were not particularly discriminable and that participants were forced to guess. While Toledo et al. (2019) and Gleize et al. (2019) employed the strict choice procedure as we did, Habernal & Gurevych (2016) gave raters the option to say that rationales were equally convincing. We suspect that this would have greatly increased reliability.

The two studies described above establish that forced choice assessments of QoR have high criterion validity and reasonable inter-rater reliability. The results provided by forced choice are consistent between expert and non-expert raters, the protocol itself requires little-to-no training, and the decisions between products can be completed within short time limits (i.e., within a minute). These findings support the use of forced choice assessments of QoR to generate a "gold standard" annotated corpus of arguments that can then be used to validate pointwise methods (Getz et al., 2020) and to train a neural model that can generate automated scores of isolated products (Habernal & Gurevych, 2016; Toledo et al., 2019). Our results also prove the method is applicable to longer written products than so far investigated (50-500 in this study compared to, e.g., 8-36 in Toledo et al., 2019). Moreover, the method is context-neutral and could be adapted to evaluate arguments about a variety of topics reliably, including ones for which there is no normatively correct solution.

Acknowledgments We would like to thank Mark Burgman and three anonymous referees for helpful comments on an earlier version of this paper.

Funding This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) (2016), under Contract (16122000002). The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the United States Government. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

Conflicts of Interest The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Ethics Approval All procedures have been approved by the Office of Research Ethics and Integrity, University of Melbourne, Protocol: Improving Reasoning through Argument Marshalling No: 11354.

Availability of Data and Materials. The datasets generated and/or analysed are available on OSF, https://osf.io/9qhxn/files/. Except for the in-house problems, all other problems analyzed in this study are available from the third parties identified in the bibliography, but restrictions may apply to their availability. Therefore, items not included in these supplementary materials can only be accessed with permission from the licensor.

Authors' Contributions. Conceptualization: SD,MW,AM,LR,AB,TP; Data curation: MW,SD; Formal Analysis: MW,SD; Funding acquisition: SD,TvG; Investigation: MW,AM,LR,BS,AK; Methodology: MW,SD; Resources: AK,MS,TvG; Software: BS,MD; Project administration: MW,AM,SD; Writing – original draft: MW,AM,SD; Writing – review & editing:AM,SD,MW,LR,AK,TvG,TP,AB,BP,MD,MS.

References

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. D. (2006). The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, 113(4), 700.

Cao, Z., Qin, T., Liu, T. Y., Tsai, M. F., & Li, H. (2007, June). Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning* (pp. 129-136).

Engel, D., Woolley, A. W., Jing, L. X., Chabris, C. F., & Malone, T. W. (2014a). Reading the Mind in the Eyes or Reading between the Lines? Theory of Mind Predicts Collective Intelligence Equally Well Online and Face-To-Face. *PLoS ONE*, *9*(12), 1–16. https://doi.org/10.1371/journal.pone.0115212

Ennis, R. H., Millman, J., & Tomko, T. N. (1985). *Cornell critical thinking tests level X & level Z: Manual*. Pacific Grove, CA: Midwest Publications.

Gleize, M., Shnarch, E., Choshen, L., Dankin, L., Moshkowich, G., Aharonov, R., & Slonim, N. (2019). Are you convinced? Choosing the more convincing evidence with a Siamese network. *arXiv preprint arXiv:1907.08971*.

GMAT Official Guide 2018. (2018). Hoboken, NJ: John Wiley & Sons, Inc.

Gretz, S., Friedman, R., Cohen-Karlik, E., Toledo, A., Lahav, D., Aharonov, R., & Slonim, N. (2020, April). A large-scale dataset for argument quality ranking: Construction and analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 05, pp. 7805-7813).

Habernal, I., & Gurevych, I. (2016, August). Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1589-1599).

Lagnado, D., Liefgreen, A., Pilditch, T. (2017). *BARD Problem Series: Spy Messaging.* (part of the Bayesian ARgumentation via Delphi, BARD problem series). Developed in partnership with University College London (London, UK), Birkbeck (London, UK), and Monash University (Melbourne, Australia)

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, 159-174.

Link, S. W. (1975). The relative judgment theory of two choice response time. *Journal of Mathematical Psychology*, 12(1), 114-135.

Mandel, D. R. (2015). Instruction in information structuring improves Bayesian judgment in intelligence analysts. *Frontiers in psychology*, *6*, 387.

Manhattan Review. (2012). *Turbocharge Your GMAT Integrated Reasoning Study Guide*. Manhattan Review Test Prep & Admissions Consulting. New York, NY

McGrath, J. E. (1984). *Groups: Interaction and performance*. Englewood Cliffs, N.J.: Prentice-Hall, c1984.

Riedl, C., Kim, Y. J., Gupta, P., Malone, T. W., & Woolley, A. W. (2021). Quantifying collective intelligence in human groups. *Proceedings of the National Academy of Sciences*, *118*(21), e2005737118. https://doi.org/10.1073/pnas.2005737118

Toledo, A., Gretz, S., Cohen-Karlik, E., Friedman, R., Venezian, E., Lahav, D., ... & Slonim, N. (2019). Automatic Argument Quality Assessment--New Datasets and Methods. *arXiv preprint arXiv:1909.01007*.

van Gelder, T., Kruger, A., Thomman, S., de Rozario, R., Silver, E., Saletta, M., ... & Burgman, M. (2020). Improving Analytic Reasoning via Crowdsourcing and Structured Analytic Techniques. *Journal of Cognitive Engineering and Decision Making*, *14*(3), 195-217.

Wood, J. (2013). *Errors of reasoning*. Studies in Logic 45. College Publications, London.

Wachsmuth, H., Naderi, N., Hou, Y., Bilu, Y., Prabhakaran, V., Thijm, T. A., ... & Stein, B. (2017). Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (pp. 176-187).

Woolley, A. W., Chabris, C. F., Pentland, A., Hashmi, N., & Malone, T. W. (2010). Evidence for a Collective Intelligence Factor in the Performance of Human Groups. *Science*, *330*(6004), 686–688. https://doi.org/10.1126/science.1193147