INTELLIGENCE AND NATIONAL SECURITY
https://doi.org/10.1080/02684527.2022.2026584

Routledge
Taylor & Francis Group

Using argument mapping to improve clarity and rigour in written intelligence products

Ariel Kruger, Luke Thorburn and Timothy van Gelder

ABSTRACT

Intelligence products should be clear and rigorous. Intelligence organisations often attempt to improve clarity and rigour with training in thinking and writing. It is assumed that (1) training enhances clarity and rigour in products, and (2) greater clarity and rigour leads to better decisions. We describe three studies exploring these assumptions. These studies concern an initiative in a biosecurity context, based on a form of argument mapping. Results indicate that reports improved, but were inconclusive about the quality of decisions. We discuss implications of this research for intelligence.

Introduction

Intelligence products are the primary means by which intelligence is provided to decision makers. For this reason, their quality is pivotal to the success of the intelligence enterprise. As Arthur Hulnick put it, 'providing intelligence to the policy system – no matter how it is collected or analysed – is in many respects the "end game" of the intelligence process, and unless the game is well played, the other moves made by intelligence to collect and analyse information might well be wasted'.[1]

Playing the game well means, in part, delivering products that are clear and rigorous. This is reflected in the US Intelligence Community (IC) Analytic Tradecraft Standards, which are broadly aimed at 'excellence, integrity, and rigor' in analytic thinking, and specifically require products to exhibit 'clear and logical argumentation'.[2] Similarly, the UK PHIA's Common Analytic Standards include the requirements 'Clear: … Analytical products should demonstrate clarity of thought by presenting clear and consistent assessments supported by coherent reasoning and relevant information' and 'Rigorous: Analysts … should show logical and coherent reasoning … '[3]

These brief excerpts highlight the intimate relationship between analytic rigour and logical reasoning. There are, certainly, other aspects of analytic rigour,[4] and analytic rigour is only one virtue of good intelligence products. In this paper, however, our focus is on logical rigour, and where we use the term 'rigour', the more cumbersome term 'logical rigour' could generally be used instead.

One way intelligence organisations have pursued clarity and rigour in products is by training analysts in thinking and writing. An example is 'Analysis 101' launched in the US IC in the mid-2000s by the Office of the Director National Intelligence. This has been described as 'a foundational course in the critical thinking, structured techniques, and other tradecraft skills that rigorous analysis demands, regardless of the functional or geographic specialization'.[5]

Two assumptions underpin these directives and efforts. One is that *training works*, in the sense that training analysts in thinking and writing boosts the clarity and rigour in the products they go on to generate. The second is that *clarity and rigour work*, or in other words, that greater clarity and rigour makes products more useful to decision makers and leads to better decisions.

CONTACT Ariel Kruger ariel.kruger@unimelb.edu.au

These assumptions seem very plausible, but they should still rigorously evaluated. If they turn out to be false, the intelligence profession would need some major rethinking. More likely, they have some truth, though not as much as we hope and imagine. Rigorous evaluation would help inform more disciplined efforts to increase clarity and rigour, and to make products more useful.

As best we can tell, little if any such evaluation has taken place. Our search of publicly available sources, such as academic journals and government websites, has not turned up any research directly addressing whether training improves clarity and rigour in real intelligence products produced by analysts who had undergone the training, or whether increased clarity and rigour in products improves the decisions informed by those products.

Of course, there is always research of indirect relevance. For example, Shelagh Dorn studied a six-week online training program in writing for intelligence.[6] Writing samples produced in the course context were assessed using an intelligence-oriented scoring rubric. Samples produced after the training scored slightly higher on the rubric than those produced before it. This is only indirectly relevant because it focuses on improvement in samples produced in writing exercises. It does not examine whether there was any 'on the job' impact on real intelligence products produced by analysts who had undergone the training, which is our concern. A second example illustrates a similar relevance gap. In a survey of analysts in one US intelligence organisation, Stephen Coulthart found some evidence that when analysts are trained in structured analytic techniques (SATs), they are more likely to use those techniques; also, the analysts tended to believe that SAT use increases rigour in products, and this belief is associated with greater use of SATs.[7] This research did not study the clarity and rigour of the products themselves, and so could not establish that SAT training had in fact increased clarity and rigour in those products. Finally, from another angle, many studies have examined the effect of critical thinking training on the development of critical thinking skills in undergraduate students. They generally find a positive effect[8]; this gives some basis for optimism about similar training for intelligence analysts, but again it does not establish any improvement in the resulting products.

Studies relating indirectly to the second assumption are harder to find, but there has been some work in the health technology area. For example, Poder and colleagues found that lack of clarity in reports making recommendations regarding adoption of health technologies was one impediment to the acceptance of those recommendations.[9]

Various factors might help explain the lack of research directly targeting these key assumptions. These include:

- Intelligence organisations have generally been slow to subject their practices to scientific scrutiny.[10]
- The two assumptions have *prima facie* plausibility. Claims which seem obviously true are often, and understandably, the last to be questioned.
- Security considerations can be a major obstacle to any such research. Studying the two assumptions means looking very closely at intelligence products and the decisions with which they were connected. Outside researchers often do not have sufficiently trusted status (e.g., clearances) and the organisations themselves usually don't have, or can't allocate, expertise in social science.
- In general, rigorously evaluating the impact of training on workplace performance is methodologically challenging and resource-intensive. This is why evaluation so often goes no further than asking about participant satisfaction using simple surveys.[11]
- Evaluating the second assumption, that clarity and rigour lead to better decisions, requires some way of assessing the quality of those decisions. However, it is very hard to assess complex, real-world decisions.[12]

This is a formidable array of challenges, but it may still be possible to make some progress. In this paper, we report on work evaluating the two assumptions somewhat more directly than prior research. We describe three exploratory studies evaluating the impact of a training-based initiative,

addressing two main research questions: (1) to what extent had the initiative improved the clarity and rigour of reasoning in written reports and (2) to what extent had this improvement led to better decisions?

The context of our work is biosecurity risk analysis in the Plants division of the New Zealand Ministry of Primary Industries (NZMPI). This setting is at one remove from intelligence, but it had some important advantages. First, we were able to work with an organisation that was happy to have such research conducted on itself. Second, the reports were generated and the relevant decisions were made within the one organisation, so we could easily identify and study both, and their relationships. Finally, security was not a barrier; we had full access to the reports and decisions, which were publicly available. These advantages were however secured at the cost of some relevance to intelligence – a topic to which we return in the final section.

NZMPI's initiative consisted largely of workshops covering a form of argument mapping. Originating in college-level critical thinking instruction, argument mapping has found a place in the repertoire of structured analytic techniques (SATs) for intelligence analysis,[13] and has been part of the curriculum in Intelligence 101.[14] It is widely assumed that using SATs (and so, presumably, training in their use) improves the quality of intelligence analysis,[15] and SAT use in the US Intelligence Community has even been mandated by Congress. However there has been little research generally into the impact of SATs,[16] and to our knowledge none on the impact of argument mapping specifically.[17] Overall, the SAT research that has been done struggles to find convincing evidence that SAT use improves analysis.[18] Thus our studies help address a major gap in our understanding of how intelligence can be improved.

The next section lays out the background to our project; the following one describes our three studies. Overall, we found that the initiative improved the clarity and rigour in written products, but we were unable to detect any impact of this improvement on decision making. In the final section, we return to explore the implications of this work for intelligence.

Background

Biosecurity risk analysis and decision making at NZMPI

NZMPI is the New Zealand government agency responsible for biosecurity. It aims to keep pests and diseases out of the country and eradicate or control those already present. This includes managing the risks posed by pests and diseases associated with imported plant and animal products, using risk mitigation measures.

Mitigation measures vary in their capacity to reduce risk. Generally, the more effective the measure, the more it restricts trade. For example, if the fruit fly b*actrocera dorsalis* is detected on pears arriving from China, the pears are either reshipped away from New Zealand or destroyed, and pear imports from China are suspended. This negates a very substantial risk, but it stops trade. However, detection of the mite *Tetranychus truncatus* only requires brushing dirt and other organic matter from the fruit. This measure does not reduce risk as much, but is appropriate because the mite poses much lower risk to New Zealand.

The measures to be applied for an imported product are specified in a legal document called an Import Health Standard (IHS). The IHS details the decisions reached as to the measures needed for an 'appropriate level of protection' from pests and diseases associated with the import. These decisions must pay regard to both the benefits of trade, and the risks to New Zealand. The assessment of these risks must have a scientific basis.[19]

Risk assessments are laid out in reports called an Import Risk Analysis (IRA). These describe the risk for each pest or disease associated with an import. The risk for a specific pest/disease depends on a number of finer-grained estimates of likelihood or significance. Each of these estimates is supported by reasoning grounded in evidence and presented in technical prose, of length typically ranging from a few paragraphs to a couple of pages. We call these sections of technical prose 'logical units'

(LUs); examples are shown in Figure 2. An IRA document might run to hundreds of pages and contain many dozen LUs, each one being of a particular type, such as a Likelihood of Entry assessment, or an Economic Consequences assessment.

Relating biosecurity risk analysis and decision making to intelligence

Biosecurity management decisions are critical to protecting the population, economy, environment and socio-cultural practices from biological threats while at the same time allowing commerce to thrive, and so a balance must be struck that protects both interests. Much the same is true of intelligence-informed decisions, at least in a democratic system. They too aim to protect the population, economy, environment, socio-cultural practices and more, while avoiding harms such as restrictions on personal liberty and intrusions on privacy. Thus, decision making in both domains involves balancing diverse considerations.

Moreover, there are some deep similarities between biosecurity risk analysis and intelligence. One way to show this is to compare the cycle of intelligence production with NZMPI's risk analysis process. The classic intelligence cycle has five components[20]: **planning and direction** whereby a consumer (decision or policy maker) requests intelligence on a specific matter; **collection**, involving obtaining raw information from a variety of sources; **processing** of raw information into a usable form; **analysis**, where information is evaluated, interpreted, and combined to generate findings addressing the consumer's needs; and **dissemination**, the delivery of the analysis to the consumer, generally in written reports and oral briefings. While the classic cycle is now regarded as a poor depiction of how intelligence actually unfolds, the five components tend to persist in alternative models, since they are essential aspects of intelligence work.[21]

In the NZMPI risk analysis process, **planning and direction** is the commissioning of an IRA to assess the pest and disease risks associated with importing a specific commodity.[22] This is similar to corresponding stage in the intelligence cycle where intelligence is requested on a specific matter. Moreover, in both the intelligence cycle and the NZMPI risk analysis process, the request is made by the same entity that will, ideally, use it to make a decision or inform policy.

Corresponding to the next two stages in the intelligence cycle, **collection** and **processing**, there is an information-gathering aspect to biosecurity risk assessment. Here NZMPI analysts seek relevant information from a range of sources such as the scientific literature, national and international databases, and outside experts. The biosecurity analysts must locate the critical needles in the vast haystacks of information. An important part of this information-gathering is identifying potential hazards associated with an import. Hazards are pests, pathogens or diseases which could be introduced into New Zealand that are capable of, or potentially capable of, causing unwanted harm.[23] Analysts develop an initial list, and further process that list by assessing each organism or disease against criteria to determine if they pose enough risk to warrant more in-depth assessment.[24]

In the **analysis** phase, biosecurity analysts build on the gathered information to develop an overall assessment of the risk associated with each hazard on the final list. This activity is conducted within a structured framework under which analysts must consider a range of critical issues and, for each one, make a judgment based on all obtainable relevant information. For example, one critical issue is the likelihood that a specific pest or disease, such as *Bactrocera dorsalis*, will *enter* New Zealand with the importation of a particular product, such as pears from China. An analyst must choose a verbal likelihood of entry and support that choice with reasoning grounded in evidence. The analyst's judgements on critical issues are then systematically aggregated by the framework into the overall risk assessment for that pest or disease. As with the analysis phase in the intelligence cycle, information (factors pertaining to pest/disease risk) has been evaluated and interpreted to develop a product (overall risk assessment) that address the consumer's needs (determining what level of protection is appropriate).

Finally, in both the biosecurity and intelligence context, the **dissemination** phase is where the completed product is delivered to the consumer. In the biosecurity context, the consumer is the decision maker who specifies which risk mitigation measures are to be imposed on an import as well as any parties to whom the risk assessment is relevant (stakeholders, other MPI departments). Likewise in the intelligence context, the primary consumer is the decision or policy maker who requested the intelligence as well as relevant other parties who 'need to know'. Ideally then, in both contexts, decision/policy makers use the completed product to inform their decisions.[25] Thus, the final phase of the both the intelligence cycle and NZ's biosecurity risk analysis process concludes with handing over a completed product to decision makers.

There are of course also many differences between intelligence and the biosecurity risk analysis. However, the fact that each stage in the traditional intelligence cycle can find a counterpart in the biosecurity risk analysis process suggests there are sufficient similarities for biosecurity to be a reasonable proxy in exploring the impact of training in thinking and writing on reports, and on the quality of resulting decisions.

The initiative

In 2015, NZMPI commenced an initiative aimed at improving the quality of IRAs, focusing on the clarity and rigour of the reasoning in their constituent logical units. Improving IRAs in this regard was expected to have various benefits.[26] One was improving the risk *judgements* reported in the IRA. Better reasoning in the logical units should lead to better judgements on those critical issues, and ultimately to better assessment of the overall risk associated with importing a particular commodity. A second expected benefit was better *decisions*. Greater clarity and rigour in the reasoning should make the risk assessments more credible and persuasive to decision makers, helping ensure that their decisions better reflect the real nature of the risks. Third, greater clarity and rigour should help *communicate* the thinking behind risk assessments and policy decisions to various other audiences, particularly stakeholders such as importers.

In what way were IRAs perceived to have insufficient clarity and rigour? In terms of clarity, readers of the IRA, decision makers and stakeholders often struggled to see how the presented information amounted to a reasoned basis for the corresponding judgements. For example, if the likelihood that a particular pest would enter New Zealand with an imported commodity was claimed to be *low*, how exactly did the mass of technical detail presented in the logical unit justify *low*, rather than *moderate* or even *high*? The logical threads may have been intuitively apparent to the experienced analysts writing the logical units but were not nearly so obvious to readers lacking their expertise and familiarity with the topic.[27]

Thus, NZMPI sought a way to ensure that the logical structure in logical units would be more transparently displayed. Such transparency would enable more rigorous scrutiny of that structure, leading (if needed) to correction of the reasoning and modification of the resulting judgements. This should lead in turn to better decisions and better stakeholder acceptance.

The transparent display of logical structure is the essence of argument mapping. At one level, argument mapping is just drawing diagrams of reasoning. These diagrams are superficially similar to other kinds of 'maps' such as mind maps, concept maps, and flowcharts, but use different layouts and visual conventions. At a deeper level, argument mapping is making explicit the logical relationships between claims, subject to rules and conventions governing how those claims should be articulated and how the relationships should be structured.[28]

It therefore made sense for NZMPI to try incorporating argument mapping into the IRA development process. They did not of course intend to replace the prose in the IRAs with diagrams. Rather, they hoped that mapping would help analysts make the logical structure of logical units more explicit, which could then guide the presentation of reasoning and evidence in the prose.

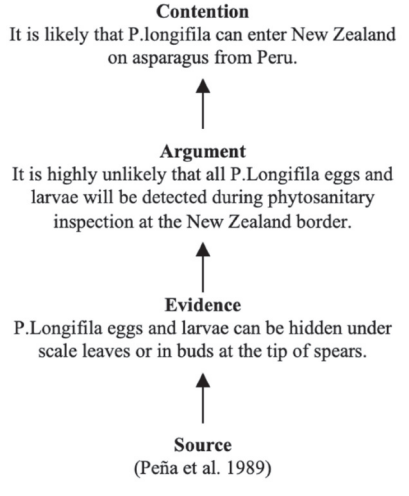
The MPI initiative had two main components. One was training in argument mapping; the other consisted of efforts to ensure that argument mapping was actually used in writing reports. The training consisted primarily of one-day workshops on argument mapping delivered by external instructors with expertise in argument mapping. Seven such workshops were offered to analysts and managers in the period 2015–2018. In addition, groups of analysts and managers held their own self-directed practice sessions, and some managers guided their teams in applying argument mapping principles in their IRA writing.

The CASE approach

The NZMPI initiative used a variety of argument mapping known as CASE. This is an acronym for Contention, Argument, Evidence, and Source,[29] the four most basic components of the CASE argument scheme. An argument scheme is a template representing a common pattern of reasoning or argumentation.[30] The core CASE scheme, as shown in Figure 1, involves a Contention (the main claim being argued for), supported by an Argument claim, backed up by an Evidence claim, obtained from some Source. This core scheme reflects the essential structure of a logical unit in an IRA, where a contention (the judgment, such as a likelihood of entry) is supported by an argument or arguments, grounded in detailed information (as described above), obtained from specifiable sources such as scientific publications.

The core CASE scheme is the kernel of a more complex, dynamic scheme with more components, complemented by a body of theory, guidelines and techniques for articulating, structuring, strengthening and presenting reasoning on almost any topic.[31] Important aspects of this framework include multiple lines of argument bearing on a contention, and multiple items of evidence supporting an argument; multiple levels of argument (sub-arguments, and so on); rules for using the principle of abstraction in articulating and organising these argument levels; 'negative' arguments and evidence (i.e., objections, and counter-evidence); the compound nature of individual arguments, and items of

(a)



(b)

[C] It is likely that P.Longifila can enter New Zealand on asparagus from Peru.

> [A] It is highly unlikely that all P.Longifila eggs and larvae will be detected during phytosanitary inspection at the New Zealand border.
>
> - [E] P.Longifila eggs and larvae can be hidden under scale leaves or in buds at the tip of spears [S] (Peña et al. 1989).

Figure 1. A simple piece of reasoning fitting the core CASE argument scheme. On the left, the reasoning is in standard argument map (i.e., diagrammatic) format. On the right, the same reasoning is presented in a form of structured writing called 'CASE-mapped prose'. CASE-mapped prose forms a logical skeleton for the final written presentation of the reasoning.

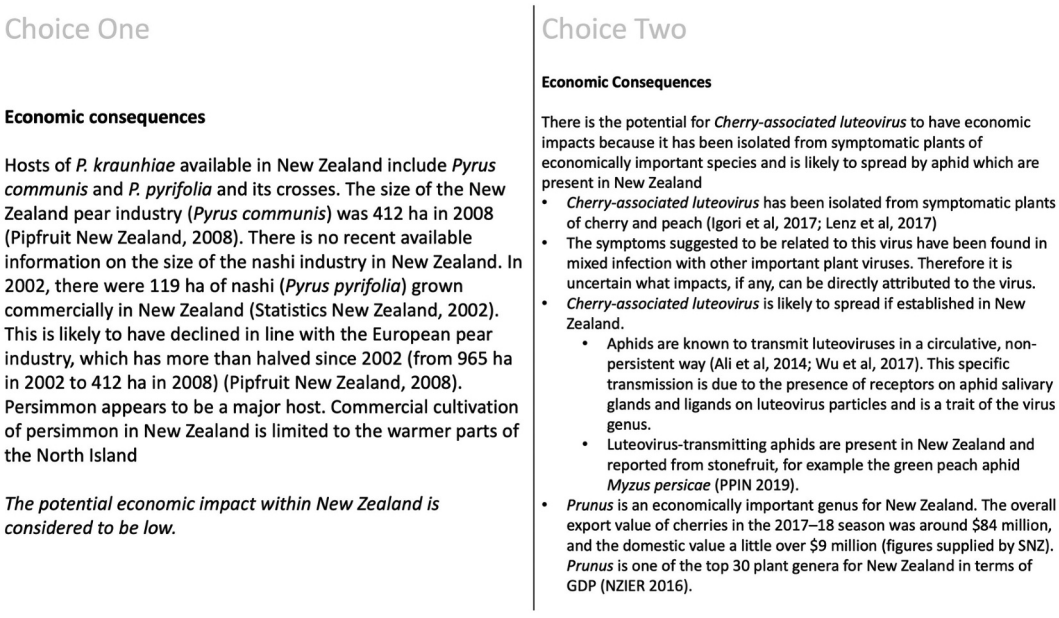


Figure 2. Illustration of the simultaneous presentation of a pre-CASE and a CASE-era logical unit, as displayed to subject in Study 2.

evidence, which both necessarily involve additional claims which in CASE are called bridging claims,[32] or (hidden) assumptions; and rules for articulating the set of claims constituting a given argument or item of evidence, which help expose hidden assumptions.

Reasoning incorporating some or all of these aspects can become quite complex. Such reasoning is often best visualised, with diagrams created and manipulated using software designed for this purpose.[33] However the same structure can also be represented in 'CASE-mapped' prose. This is a form of structured writing in which the logical structure presented diagrammatically in an argument map is presented in prose using a set of formatting conventions, as illustrated in Figure 1. These conventions include having the Contention stated at the top (i.e., 'bottom line up front') and using bullet-points for one purpose only – to indicate that the bulleted lines are items of Evidence bearing directly on the claim under which they are nested. More generally, CASE-mapped prose adheres to the principle that the reasoning bearing upon any claim in the logical structure is nested directly underneath that claim. Thus, to know what parts of the overall text are intended to have some logical bearing on a given claim, you should – and need only – look to what is nested under that claim. Conversely, you can easily ascertain the logical role of any claim, for it is to provide logical support, or opposition, to the claim 'above' it in the indentation hierarchy.

CASE-mapped prose can be used as an outline when writing a logical unit in an IRA. The austerity of CASE-mapped prose can, and generally should, be softened by massaging the writing to be more fluidly readable, though in doing this care is needed to avoid losing transparency of logical structure.

Each one-day training workshop at NZMPI introduced a cohort of participants to the concepts and techniques of CASE mapping as just sketched. Development of understanding and skill was driven by numerous exercises of increasing difficulty, with most exercises drawn from the biosecurity domain, using examples from real work in NZMPI or similar organisations.

Evaluating the initiative

The CASE initiative was a substantial investment of resources by NZMPI. While there was no organisational requirement to use CASE, we know anecdotally that analyst teams would meet regularly to discuss how to present their findings in CASE format. Analysts who took part in the initiative kept their CASE instructional material.

The case for CASE might have been compelling in theory, but the Ministry wanted to know whether it was working in practice. For help, NZMPI turned to the Centre for Excellence in Biosecurity Risk Analysis at the University of Melbourne (CEBRA), with which it has collaborated for many years, leading to the current project.

We separated the issue of whether the initiative was working into two main research questions. The first asked whether the initiative was making a difference to *reports*. To what extent had the CASE initiative led to greater clarity and rigour in their reasoning? To explore this, we conducted two studies. Both focused on differences between IRAs produced before, and after, the CASE initiative had commenced. We refer to these periods as 'pre-CASE' and 'CASE-era', because the CASE initiative was still unfolding during CASE-era period. One of these studies asked whether CASE-era IRAs showed stronger CASE structure than pre-CASE IRAs; the other asked whether CASE-era IRAs were better reasoned in the eyes of independent readers.

The second research question asked whether the CASE initiative was making a difference to *decisions*. To what extent were CASE-era decisions better? We conducted one study exploring this. It asked whether IHS decisions informed by CASE-era IRAs were better *aligned* with the risk assessments in those IRAs than IHS decisions informed by pre-CASE IRAs. As explained below, alignment is one dimension of overall decision quality; it is whether the severity of measures imposed by an IHS decision are commensurate with the level of risk.

For various reasons, the three studies were exploratory in nature. We did not register any specific hypotheses regarding anticipated differences between pre- and CASE-era reports or decisions. The project was as much about developing and testing methods for addressing questions as it was about answering those questions. The quantity of data we could gather, and in some respects our methods, were limited by incidental factors. Since the CASE initiative took place primarily in the Plants division of NZMPI, our research focused on reports and decisions from that division, and so our findings are limited to that context.

The next three subsections provide succinct overviews of our three studies; more detailed accounts can be found in our technical report.[34] The subsequent subsection discusses implications for our two research questions.

Study 1 – Do CASE-era reports show stronger CASE structure?

Our first study sought to determine to what extent CASE structure had been adopted in IRAs produced subsequent to the training.

Question

Do CASE-era IRAs have stronger CASE structure than pre-CASE IRAs?

Design

Study 1 used a retrospective observational design. Since it was not feasible to examine all IRAs in their entirety, we drew a sample of pre- and CASE-era IRAs, and sampled logical units (LUs) from within those reports. We developed and applied a scheme for coding these LUs for the presence of CASE structure, generating a CASE score for each one. We used descriptive statistics and mixed effects modelling to understand the difference between the pre- and CASE-era samples, and cautiously generalised to all Plant division IRAs in the relevant period.

Sampling

As just indicated, our sampling operated at two levels. First, with the help of a NZMPI expert, we carefully selected a small number of pre- and CASE-era IRAs. We had four reasons for using purposive, rather than random, sampling. First, we wanted to ensure that the sample reflected all Plant division IRAs in some key respects, such as covering the most common types of IRAs produced in that division. Second, we wanted to ensure that the selected CASE-era IRAs were well matched with the selected pre-CASE IRAs, so as to minimize confounding by irrelevant differences (such as the type of commodity they analysed e.g., fruit/vegetable, nursery stock, machinery etc.). Third, looking ahead to our third study, we wanted IRAs for which there was a corresponding IHS. Finally, as an exploratory study, we wanted to ensure we had a good chance of identifying a pre-post difference in CASE structure if there was one, so we chose to ensure that a particular CASE-era IRA was included. This IRA was developed with an explicit effort to ensure that CASE structure was well-instantiated. Our deliberate inclusion of this IRA amounted to a selection bias, which would need to be taken into account when making inferences from our results.

The final selection consisted of two pre- and two CASE-era IRAs, with both sub-samples including a fresh fruit IRA and a nursery stock IRA. The second level of sampling was to select logical units from these reports. Here, we randomly selected up to 25 logical units from each IRA, subject to some constraints aimed at maximising representativeness, and suitable matching in the pre- and CASE-era sub-samples. The final selection consisted of 92 LUs, 50 pre-CASE and 42 CASE-era.[35]

Procedure

We developed the coding scheme in three stages. First, we developed a list of nine features which ought to be identifiable in a logical unit if it transparently displayed logical structure as specified by the CASE approach. For example, one feature was the presence of a main contention (i.e., a clear point being argued for). Features were treated as binary (present or not) and a logical unit was given one point for each feature. Second, we pilot-tested the scheme on a selection of 25 logical units randomly selected from our four IRAs. It became apparent that the CASE features were often only partially present. Thus, in the third stage, we revised the scheme to allow partial credit.

We then applied the coding scheme to all logical units in our samples. This required expertise in identifying logical structure in technical prose, and a strong grasp of CASE. Since this combination is not common, all coding was done by a single expert, who was not 'blind' as to whether a logical unit was pre- or CASE-era. For each logical unit, we recorded a score for each question, resulting in an overall score out of nine.

Results

Out of a possible 9 points, the mean overall CASE score for logical units was 3.05 (95% CI 2.58, 3.51) in pre-CASE IRAs and 7.94 (7.61, 8.27) in CASE-era IRAs. The effect size for this difference (Cohen's *d*)[36] is 1.73, which is very large according to Cohen's widely used conventions.[37]

This substantial difference might have been due, in part or whole, to differences in the IRAs other than being pre- or CASE-era. To assess this, we used a linear mixed-effects regression, modelling the CASE score for each logical unit as a function of whether the source IRA was pre- or CASE-era (fixed effect), whether the IRA was for plants or produce (fixed), the type of logical unit (e.g., Likelihood of Entry) (fixed), and the IRA identity (random). Whether the source IRA was pre- or CASE-era continued to have a large positive effect. Specifically, in the presence of other explanatory variables, mean CASE-era scores were estimated to be higher than pre-CASE scores by 6.81 points (4.37, 9.26[38]).

Beyond overall scores, we were interested in specific respects in which pre- and CASE-era reasoning differed. Unsurprisingly, pre-CASE reports failed to exhibit those CASE features which amount to relatively superficial formatting requirements, such as putting the contention at the top, and using bullet points only for items of evidence. However, they also tended to lack more fundamental aspects

of clear and rigorous reasoning, such as identifiable high-level argument structure connecting the evidence to the contention. This is consistent with the prior informal observation, mentioned above, that logical units often appeared to be little more than evidence dumps.

Discussion

These results show two things about the IRAs in our sample. First, the CASE-era IRAs transparently displayed their basic logical structure. Since the maximum score a logical unit could achieve on our coding scheme was nine, a mean score of nearly eight suggests that CASE-era IRAs conformed to the CASE structure rather well. Second, the CASE-era IRAs were much better in this regard than their counterparts from before the initiative.

The results are likely to be exaggerated by observer bias. The coder was not fully independent of the CASE initiative,[39] or blind as to whether LUs were pre- or CASE-era. These factors may have caused the coder to see stronger CASE structure in CASE-era IRAs, despite their being aware of the potential for bias and attempting to avoid it. However, the difference between the pre and CASE-era LUs was dramatic and we find it implausible that observer bias accounts for the bulk of that difference.

We cautiously infer that, more generally, CASE-era Plant division IRAs show much stronger CASE than their pre-CASE counterparts. This inference is tempered by the observer bias, just mentioned; the selection bias described earlier, where we deliberately including a CASE-era IRA most likely to show strong CASE structure; by the possibility of random variation, given we only examined two of the dozens of pre-CASE IRAs; and our lack of detailed data on the extent to which the analysts responsible for the CASE-era IRAs had participated in CASE training. For these reasons the true difference is probably not as dramatic as found in our sample.

Study 2 – Are CASE-era reports better reasoned?

Study 1 found evidence that CASE-era IRAs exhibit stronger CASE structure than pre-CASE IRAs, as judged by a researcher with relevant expertise. This suggests that CASE-era IRAs do have greater clarity and rigour. However, the 'consumers' of IRAs, such MPI policy makers and external stakeholders, generally do not have specialist expertise in structured argumentation, and are not trained in CASE. Study 2 therefore investigates whether the difference between pre- and CASE-era IRAs is recognised, by a much wider class of readers, as better reasoning. It also addresses the observer bias in Study 1 by obtaining evidence from independent assessors.

Question

Are CASE-era IRAs better reasoned, as judged by general readers?

Design

Like Study 1, this was a retrospective observational study. We recruited cloud workers to be our sample of general readers, and drew a sample of LUs from both pre- and CASE-era IRAs. Participants ('subjects' in what follows) were asked to evaluate the quality of reasoning in the LUs. We used descriptive statistics and regression analysis to estimate the difference in quality between pre- and CASE-era LUs.

We obtained subjects' evaluations using 'two alternative forced choice' (2AFC). This is an experimental paradigm in psychophysics[40] and cognitive psychology,[41] in which two stimuli are presented simultaneously, and the subject must select one on some criterion.[42] For example, two colours are presented, and the subject must choose the one they think is brighter. 2AFC has previously been used to evaluate reasoning. For example, it has been used to test subjects' ability to recognise deductive validity.[43] A recent study tested 2AFC as an alternative to more traditional methods for evaluating the quality of reasoning in short passages comparable to the LUs in this project. It found that untrained subjects using 2AFC largely agreed with trained evaluators using a complex scoring rubric, the Intelligence Community Rating Scale,[44] despite using much less time and effort.[45]

Sampling

We obtained a convenience sample[46] of general readers by recruiting cloud workers on Amazon Mechanical Turk. This is a crowdsourcing platform on which individuals can sign up for paid participation in a wide range of online tasks including studies requiring human subjects.[47] Opportunities to partake in such studies are posted on the platform, and interested 'Turkers' can elect to be involved. These Turkers are quite diverse demographically. We recruited 89 Turkers.[48]

As in Study 1, we obtained a purposive sample of LUs in two stages. First, we selected pre- and CASE-era IRAs. In this case, for CASE-era IRAs, we chose to use only the CASE-era IRA found in Study 1 to have the strongest CASE structure. We paired that with a pre-CASE IRA of the same type (nursery stock). Second, from each of these IRAs, we randomly selected 58 LUs, 29 from each of two LU types.[49] The mean CASE scores were 2.73 (pre) and 8.45 (post). Thus, subjects would be forced to choose between LUs which, on average, differed markedly in the extent to which they exhibited CASE structure.

Procedure

Subjects were first asked to study a short training module containing general instructions and simple guidance for making selections. They then participated in a series of trials, in which they were simultaneously presented with two LUs, positioned side by side on their screens (Figure 2). One LU was randomly drawn from the pre-CASE IRA, and one from the CASE-era IRA. The CASE-era LU was randomly assigned to either the left or the right. Subjects selected the one that was 'better reasoned and communicated' by clicking on their selection. To help ensure that they were making thoughtful choices, subjects were required to provide a brief rationale for each selection. Subjects completed up to nine trials in immediate succession.

Results

The trials resulted in 765 usable selections, or about 8.5 per participant, of which 449 were the logical unit drawn from the CASE-era IRA. That is, CASE-era passages were chosen 59% of the time. The effect size (Cohen's *d*) for this difference is 0.31,[50] which is a small-medium effect.

The raw selection proportions might be confounded by variability among the subjects who made the selections. To account for potential subject effects, we used binomial (logit) generalised linear mixed-effects regression, modelling an indicator variable for whether the CASE-era logical unit was selected as a function of a constant intercept term (fixed effect), and the subject making the selection (random effect). Controlling for subject effects, we estimated the underlying probability that CASE-era logical units were thought to be better reasoned and communicated as 0.60 (0.54, 0.67). The fact that this probability is so close to the raw proportion (0.59) indicates that the net impact of subject effects is negligible.

Discussion

At face value, the data show a small but statistically strong tendency for subjects to regard the CASE-era LUs in our sample as better reasoned and communicated than their pre-CASE counterparts. However, there are reasons to be cautious about this interpretation.

First, when tending to select LUs from CASE-era IRAs, subjects might have been picking up superficial indicators, such as greater use of indenting, rather than any deeper difference in quality. To the extent that this is true, the data would be exaggerating the true quality difference.

Second, subjects' selections might have been biased by the instructions. They were told to select the LU which was 'better reasoned and communicated'. This wording was chosen for succinctness, but it might have been misleading. It seems to treat reasoning and communication as two separate aspects of the LUs. However, we were really concerned only with the reasoning: how rigorous and clearly presented was it? Participants might have judged many of the pre-CASE LUs as better

communicated in other, non-relevant ways, such as being more 'readable', even if the reasoning in those LUs was not as rigorous or clear. To the extent that this was true, it might mean our data is understating the true difference in reasoning quality.

It is difficult to estimate the net effect of these factors, because the impact of each is not quantified, and they may counter-balance. A reasonable position is that the CASE-era LUs in our sample, with their stronger CASE structure, were slightly better-reasoned in the eyes of our general readers, but the precise extent is unclear.

What can we conclude, from this finding, about all Plants division IRAs? It is tempting to infer that CASE-era IRAs more generally are slightly better reasoned in the eyes of general readers. However, this is only weakly supported by the data, given the small and purposive nature of our sample of LUs, and potential differences between our subjects (Turkers) and the larger population of general readers. We can however be moderately confident that, *to the extent that* CASE-era IRAs have stronger CASE structure, they will be perceived as slightly better reasoned by general readers.

Study 3 – Are CASE-era decisions better aligned?

The first two studies suggested some improvement in reports coincident with the CASE initiative. With our third study, our focus shifted to the decisions informed by those reports. Were they also better CASE-era? To investigate this, we needed some measure of how good those decisions were. However, it is very hard to measure the quality of complex real-world decisions. Indeed, what quality consists in is not a settled matter; there are many conflicting perspectives on decision quality (e.g., process-focused vs outcome-focused) and no obvious way to reconcile or choose between them.[51] At a practical level, as outside researchers, we did not have the domain expertise, or sufficient information, to stand in judgment on NZMPI decision making.

We therefore had to develop and apply a limited-purpose criterion of decision quality, one we could apply in a relatively objective and mechanical manner. We focused on the legislative requirement that risk management decisions be informed by the relevant risk assessments.[52] We took this to imply that decisions should generally *align* with the risk assessments: the greater the risk, the more severe the measures which should be required. We couldn't expect this alignment to be perfect; IHS decisions must also be informed by other considerations such as economic impact of those measures. Consequently, a decision maker might reasonably, on occasion, decide to apply measures which seem less, or more, severe than appropriate considering only the level of risk. Looking at sets of decisions, however, we should see broad alignment. This means that, other things being equal, a decision that is aligned with the relevant risk assessment is better than one that is not. Further, as described below, alignment can be assessed with straightforward procedures requiring little expertise-based judgment.

Question

Are CASE-era IHS decisions better-aligned with the corresponding risk assessments?

Design

As in our first two studies, our design was retrospective and observational. We obtained a sample of IRA-IHS pairs from before the start of the CASE initiative, and one from after the start of the initiative. We calculated the degree of alignment in each sample as the proportion of pests or diseases considered in the IRA-IHS pairs for which decisions and assessments were (mis)aligned. We used descriptive statistics and a regression model to assess the difference in alignment between the pre- and CASE-era samples.

Sampling

Our samples consisted of the pre- and CASE-era IRAs used in Study 1, paired with their corresponding IHSs. We thus had two pre-CASE IRA-IHS pairs and two CASE-era pairs, matched for commodity type. That is, both the pre- and CASE-era samples included one pair dealing with a fresh fruit and one dealing with a nursery stock. We examined every decision in the IHSs for which there was a corresponding risk assessment in the relevant IRA.

Procedure

We calculated alignment in three steps, but the nature of these steps differed for the two commodity types, due to a difference in the IRAs for these types. IRAs of both types assessed the risk associated with each pest or disease, but nursery stock IRAs also provided recommendations for risk mitigation measures.

For nursery stock IRA-IHS pairs, we first recorded the measure recommended in the IRA for each pest or disease. Second, we recorded the measure actually imposed by the IHS. Third, we compared the two measures. If they were the same, we treated the decision as aligned with the risk assessment.

Calculating alignment for fresh fruit IRA-IHS pairs was more complex. The IRAs in these pairs did not provide recommendations, so we couldn't just check if the recommended and applied measures were the same. We thus had to devise another way of assessing the match between the assessed risk and the applied measure. In the first step, we applied a coding scheme to the material presented in the IRA to produce an overall risk score for each pest or disease. Second, we coded the severity of the risk management measures specified for the pest or disease in the IHS. Third, we assessed whether the severity score for pest or disease was consistent with the severity scores for other pests or diseases with the same or similar risk scores. Put another way, we looked for anomalies: cases where the severity of specified measures were out of line with severity of measures applied to other equally risky pests or diseases.

Results

90 of the 101 pre-CASE decisions were aligned (89.1%) while 61 of the 65 CASE-era decisions were aligned (93.8%). Put another way, the proportion of decisions that were aligned in our samples gained 4.74% (−11.3%, 17.7%) after the start of the CASE initiative.[53] The effect size (Cohen's *d*) of 0.16 is lower than the conventional threshold for a small effect size (0.2).

To further help assess the significance of this increase, we used a binomial (logit) generalised linear regression, modelling the probability that the treatment of a given pest is aligned as a function of (1) whether the pest was considered in pre- or CASE-era IRA-IHS pair, and (2) what type of commodity the pest or disease was associated with. Our model found that the odds of alignment CASE-era was 2.15 (0.681, 8.223) times the pre-CASE odds. Note that the lower confidence bound is less than one, the point at which odds are equal, so the difference is technically not statistically significant.

Discussion

CASE-era decisions in our samples were slightly better-aligned with risk assessments than their pre-CASE counterparts. However, due to the small and statistically marginal nature of this difference and the purposive nature of our sample, we can't be confident that such a difference exists across all Plants division decisions.

It is worth noting that alignment was very strong, even in pre-CASE decisions, where nine out of every ten decisions matched the level of risk. This accords with the views of NZMPI staff we spoke to about this, based on their experience. This convergence of evidence makes it likely that alignment is generally high across all Plants division decisions. This might partly explain why there was only a marginal gain in alignment. Simply put, there wasn't much room for improvement.

Interestingly, the strong alignment suggests that the relative lack of clarity and rigour in the reasoning behind the risk assessments pre-CASE IRAs was not much of a problem for NZMPI decision makers. This may be because those decision makers, like the analysts producing the IRA reports, have strong domain expertise enabling them to 'read in' the logical structure which was not fully explicit, and which other readers would have more difficulty discerning.

The stronger CASE structure in CASE-era reports might have had a different kind of benefit for decision makers. Even if the quality of their decisions was unchanged, they might have been able to reach those decisions more quickly and easily, with flow-on benefits for their performance on other tasks. This possibility could be addressed in future research involving the decision makers themselves.

Implications for our research questions

Our first main research question was the extent to which the CASE initiative led to greater clarity and rigour in reasoning in NZMPI Plant division reports. Our first two studies both point towards greater clarity and rigour, though from different angles and with different strengths. The first provided solid support for CASE-era reports having considerably stronger CASE structure than pre-CASE reports, and thus greater logical transparency. The second indicated that this greater logical transparency tends to be seen by independent general readers as somewhat better reasoning and communication.

Did the CASE initiative *cause* this gain in clarity and rigour? Our exploratory studies were retrospective and observational in design, using small, purposeful samples. They were unable to control for confounding causal factors. We nevertheless conjecture that the gain in clarity and rigour was in fact largely caused by the CASE initiative, for two reasons. First, the gain was exactly the kind of impact that the CASE initiative was intended and designed to have. Second, while we can imagine potential confounding factors, we currently have no positive reason to think that any such factors were operating to an extent sufficient to account for all or most of the gain. For example, staff turnover at NZMPI over the relevant period may have resulted in the staff who produced the CASE-era reports having, on average, greater writing and reasoning aptitude than those producing the pre-CASE reports. We acknowledge this as a logical possibility but have no evidence that it is true.

Our second main research question was extent to which the CASE initiative led to better decisions. Our studies were inconclusive on this matter. Only our third study was directly relevant, and while it found a slight CASE-era improvement in the sampled decisions, the difference was too marginal to support a generalization to all Plants decisions. For this reason, plus the inability of the study design to rule out confounding factors, we can't be confident that the CASE intervention had any causal impact on decision quality. Of course, we are not concluding that it had no such impact.

In short, our studies provide good evidence that the CASE initiative at NZMPI appreciably improved clarity and rigour in IRA reports, but are inconclusive with respect to any impact on IHS decisions.

Implications for intelligence

In the introduction we highlighted two assumptions underpinning the training programs many intelligence organisations provide in pursuit of greater clarity and rigour in their products. We now revisit those assumptions in the light of our research on the NZMPI CASE initiative.

The first assumption, *training works*, was that training in thinking and writing results in greater clarity and rigour in products. Our first two studies support a limited version of this assumption, viz., an organisation *can* increase clarity and rigour by means of a suitable training-based initiative. Such improvement occurred in the NZMPI context, and while there are differences between biosecurity and intelligence, we see no compelling reason why an intelligence organisation might not achieve similar improvements by similar means.

Our studies do not of course imply that all current training programs in thinking and writing are achieving worthwhile gains in clarity and rigour. Much will depend on the nature of that training and its implementation. Reflecting on the NZMPI initiative suggests at least three factors contributing to success. One is that the training content should be tightly connected to the desired result. NZMPI sought to improve the logical transparency of their products; CASE is a framework for logical transparency. The training content was also tied to the intended outcome through use of examples drawn from the biosecurity risk domain, including from the work of NZMPI itself.

Second, the nature and content of the training should have solid scientific support. There is now considerable evidence that training in argument mapping improves generic critical thinking skills.[54] There is also evidence that it improves clarity and rigour in written work. A recent study found that the essays of philosophy students who had undertaken argument mapping training were markedly better than those of a control group. The trainees had '(a) structured their essays more effectively, (b) presented the arguments more accurately, and (c) better understood the relevant arguments'.[55]

Third, the training should be complemented by a strong commitment to application in analysts' work. The NZMPI initiative paired training with deliberate efforts, by at least some analysts and managers, to draft the logical units in IRAs according to CASE principles.

The second assumption was that *clarity and rigour work*, in the sense that greater clarity and rigour in intelligence products lead to better decisions by consumers of those products. Our studies neither support, nor undermine, this assumption. This is partly because we only managed to uncover a slight, statistically marginal impact of increased clarity and rigour, and only on one aspect of decision quality. It is also partly due to an important dissimilarity between biosecurity decision making at NZMPI and decision making informed by intelligence products. At NZMPI the connection between analysis and decision making is tighter than it generally is in the intelligence arena. For example, at NZMPI both occur within the one organisation within a coherent legislative and bureaucratic framework. Intelligence products are often provided to decision makers outside the organisation who have comparatively greater freedom to use those products as they see fit, and so the relationship between products and decisions is likely to be noisier.

Indeed, we suggest that our method is a more useful contribution to evaluating the second assumption than our results. That is, future research on the connection between clarity and rigour in intelligence products, and the quality of decisions informed by those products, might also render this issue tractable by focusing on the alignment between the probabilistic assessments in those products, and the decisions to which they are relevant. Such research could, for example, examine the alignment between the levels of risk assigned to terrorism suspects in intelligence reports, and the severity or urgency of responses required by decision makers. As in biosecurity, we shouldn't expect perfect alignment, but we might reasonably expect greater alignment when reports are clearer and more rigorous. If there is such a difference, it should emerge if sufficiently large numbers of reports and decisions are analysed with enough care.

Another issue we raised in the introduction was the potential relevance of our research to the wider program of evaluating structured analytic techniques (SATs). Our first two studies provide some evidence that *using* a particular SAT, argument mapping, can help improve the quality of analytic outputs. This is noteworthy for several reasons. First, it is unusual for scientific studies of SATs to deliver positive verdicts. For example, studies of the most well-known SAT, the Analysis of Competing Hypotheses (ACH), have had mixed results but increasingly fail to reveal a net benefit arising from its use.[56]

Second, it is rare for scientific studies of SATs to examine performance in real analytic work.[57] For practical reasons, studies tend to use concocted exercises differing from real work in important ways, such as being less complex, being artificially 'neat' or well-defined, and lacking the stress arising from potential real-world consequences. By contrast, our studies found improvements in the core business of NZMPI analysts and managers in the Plants division.

Third, our studies belong to growing body of evidence supporting the effectiveness of argument mapping. We mentioned above the numerous studies of argument mapping training in education. Our studies venture beyond that context, examining the impact of using argument mapping, considered as a SAT, on the quality of real analytic work. The history of positive results raises our confidence that the benefits revealed in our studies are genuine and largely attributable to the SAT, rather than incidental features of the NZMPI initiative and context.

One implication is that the intelligence community should put greater effort into exploring the potential for argument mapping to enhance analytic work. Studies similar to ours, but in a true intelligence context, could deepen our understanding of the impact of training in argument mapping, or use of argument mapping as a SAT, on the quality of intelligence products. Ideally such studies would be stronger methodologically, and better resourced, than our exploratory work. They should also have wider scope. It is not enough to ask whether argument mapping works, or the extent to which it works. Equally important questions include how the technique can be better designed or adapted to fit various kinds of intelligence work; in what intelligence contexts is it most useful; what kinds of tools are needed to help analysts use it easily and effectively; how much training, and of what kind, is needed; how artificial intelligence could augment analysts' mapping capabilities[58]; and why individuals or organisations might hesitate to adopt argument mapping, and how such hesitancy can be dispelled.

Notes

1. Hulnick, "The Intelligence Producer – Policy Consumer Linkage: A Theoretical Approach," 212.
2. ODNI, "Intelligence Community Directive (ICD) 203, Analytic Standards."
3. Professional Head of Intelligence Assessment, "Professional Development Framework: For All Source Intelligence Assessment," 12.
4. Barnett et al., "Analytic Rigour in Intelligence."
5. Immerman, "Transforming Intelligence Analysis: 'The Tail That Wags the Dog'," 86.
6. Dorn, "Teaching Intelligence Analysis Writing Skills: A Program Evaluation."
7. Coulthart, "Why Do Analysts Use Structured Analytic Techniques? An in-Depth Study of an American Intelligence Agency."
8. Abrami et al., "Strategies for Teaching Students to Think Critically: A Meta-Analysis."
9. Poder et al., "Impact of Health Technology Assessment Reports on Hospital Decision Makers–10-Year Insight from a Hospital Unit in Sherbrooke, Canada: Impact of Health Technology Assessment on Hospital Decisions."
10. Mandel, "The Occasional Maverick of Analytic Tradecraft."
11. Pineda, "Evaluation of Training in Organisations."
12. Keren and de Bruin, "On the Assessment of Decision Quality."
13. Heuer and Pherson, *Structured Techniques for Intelligence Analysis*, 213–16. The first (2007) version of ICD 203 mentions argument mapping as one example of alternative analysis, and requires that alternative analysis be incorporated where appropriate. ODNI, "Intelligence Community Directive (ICD) 203, Analytic Standards."
14. Immerman, "Transforming Analysis: The Intelligence Community's Best Kept Secret," 170.
15. In a survey of 65 experts, 81% agreed that use of SATs enhances analytic rigour. Barnett et al., "Analytic Rigour in Intelligencem," 87.
16. Chang et al., "Restructuring Structured Analytic Techniques in Intelligence"; and Marrin, "Intelligence Analysis: Structured Methods or Intuition?"
17. In research funded by the Intelligence Advanced Research Projects Activity, Thomason and colleagues conducted a series of studies investigating the impact of a combination of pedagogical techniques including argument mapping on the development of critical thinking skills in undergraduate students. Thomason, "A012. Critical Thinking Final Report." While they found positive results, this research is only indirectly relevant to the impact of argument mapping on the quality of actual intelligence work.
18. This is not the place to attempt a comprehensive survey of the literature on the controversial topic of the efficacy of SATs. We can however illustrate our claim that research on SATs struggles to find evidence of positive impact with a brief survey of some of the literature on the most well-known SAT, the Analysis of Competing Hypotheses (ACH). A review by Coulthart found some support for the efficacy of ACH (Coulthart, "An Evidence-Based Evaluation of 12 Core Structured Analytic Techniques.") and Stromer-Galley et.al. suggested that a modified version of ACH yielded a slight improvement relative to a control (Stromer-Galley et al., "Flexible versus Structured Support for Reasoning: Enhancing Analytical Reasoning through a Flexible Analytic Technique.").

However studies by others find no benefit from ACH use Whitesmith, "The Efficacy of ACH in Mitigating Serial Position Effects and Confirmation Bias in an Intelligence Analysis Scenario." or even detrimental effects Dhami, Belton, and Mandel, "The "Analysis of Competing Hypotheses" in Intelligence Analysis."

19. MAF Biosecurity NZ, "Balance in Trade," 5.
20. Phythian and Gill, "*From Intelligence Cycle to Web of Intelligence: Complexity and the Conceptualisation of Intelligence*.", 17.
21. For example, the five components are found in the Gill & Phythian "Web of Intelligence" model in Phythian and Gill, "From Intelligence Cycle to Web of Intelligence: Complexity and the Conceptualisation of Intelligence." and the Joint Doctrine Publication 2–00 model in Davies, Gustafson, and Rigdin, "The Intelligence Cycle Is Dead, Long Live the Intelligence Cycle: Rethinking Intelligence Fundamentals for a New Intelligence Doctrine."
22. Ministry for Primary Industries New Zealand, "Import Risk Analysis."
23. MAF Biosecurity NZ, "Import Risk Analysis: Process Overview," 5.
24. MAF Biosecurity NZ, 5.
25. We use "ideally" here because, as Marrin notes: "Intelligence analysis is regularly ignored by decision-makers, and frequently has limited to no impact on the decisions they make" Marrin, "Evaluating the Quality of Intelligence Analysis: By What (Mis) Measure?"
26. Kruger, van Gelder, and Thorburn, "The Impact of Evidence on Decision Making – Final Report."
27. Even if decision makers were at one point analysts, it does not immediately follow that they would have the necessary expertise or experience to follow a poorly written IRA

    This is known as the 'curse of knowledge.' Warren et al., "Marketing Ideas."
28. van Gelder, "Argument Mapping."
29. Strictly speaking, the acronym would be CAES, but switching the E and S makes the acronym more readable, meaningful and memorable. The two versions are pronounced the same way.
30. Walton, Reed, and Macagno, *Argumentation Schemes*.
31. Some of this material is covered in Davies, Barnett, and van Gelder, "Using Computer-Aided Argument Mapping to Teach Reasoning."
32. Bridging claims are similar to warrants in the well-known Toulmin argument scheme. See Toulmin, *The Uses of Argument*.
33. See van Gelder, "The Rationale for Rationale™." for description of one such software package. For the CASE workshops a custom tool was developed. See van Gelder, *The Reasoning PowerPoint App*.
34. Kruger, van Gelder, and Thorburn, "The Impact of Evidence on Decision Making – Final Report." Note that in the report we use 'component argument' rather than 'logical unit.'
35. One CASE-era IRA was smaller than the others; we coded all CAs from this report. This is why there were fewer CASE-era LUs in our sample.
36. Cohen's *d* here is calculated as the difference in means divided by the pooled standard deviation.
37. Cohen, *Statistical Power Analysis for the Behavioral Sciences*.
38. We note that the upper bound on the confidence interval of 9.26 exceeds the maximum CASE score of 9. This of course is because the confidence interval calculation assumes no upper bound on scores. The confidence interval represents the statistical uncertainty associated with our point estimate.
39. While the coder (Kruger) was not involved in the CASE initiative, he is a colleague of the person who delivered the CASE workshops (van Gelder).
40. Fechner, Howes, and Boring, *Elements of Psychophysics*.
41. Macmillan and Creelman, *Detection Theory: A User's Guide*.
42. For example, two images are presented, and the subject must choose which is brighter
43. Dube, Rotello, Caren, and Heit, "Assessing the Belief Bias Effect with ROCs: It's a Response Bias Effect."; Trippas, Handley, and Verde, "Fluency and Belief Bias in Deductive Reasoning: New Indices for Old Effects."
44. ODNI, *Rating Scale for Evaluating Analytic Tradecraft Standards with Amplified Guidance for Evaluators (Last Revised on 6 November 2015)*.
45. van Gelder, "SWARM Project Final Report."
46. Battaglia, "Convenience Sampling."
47. Paolacci, Chandler, and Ipeirotis, "Running Experiments on Amazon Mechanical Turk."
48. Subjects were paid US$10 per hour. Eligibility criteria were minimal. Participants had to be in the US, have previously participated in 100 or more similar tasks, and have a 95% completion rate for previous tasks. No attention checks were used given the reliability of MTurkers who met our criteria (95% completion rate and over 100 or more similar tasks)
49. The two CA types were Establishment Assessments, and Assessments of Economic Consequences.
50. See Equation 11 in Stanislaw and Todorov, "Calculation of Signal Detection Theory Measures." for the details of the effect size calculation.
51. See above 12.
52. Biosecurity Act 1993 No 95.

53. We calculated gain as the estimated difference in binomial proportions for the pre- and CASE-era samples. The 95% confidence interval is described here Zieliński, "A New Exact Confidence Interval for the Difference of Two Binomial Proportions."
54. van Gelder, Bissett, and Cumming, "Cultivating Expertise in Informal Reasoning"; Thomason, "A012. Critical Thinking Final Report"; Alvarez, "Does Philosophy Improve Reasoning Skills?"
55. Cullen et al., "Improving Analytical Reasoning and Argument Understanding."
56. Dhami, Belton, and Mandel, "The "Analysis of Competing Hypotheses" in Intelligence Analysis"; Whitesmith, "The Efficacy of ACH in Mitigating Serial Position Effects and Confirmation Bias in an Intelligence Analysis Scenario."
57. "Ideally, evaluation projects need to capture the impact of training and education on the *learner, their behaviour (in the workplace)*, and ultimately, to the extent that this is possible, how learning enhances national security and public safety." Walsh, "Teaching Intelligence in the Twenty-First Century." 1014; italics in original.
58. For example, Luke Thorburn has been developing a prototype 'argument processor' for AI-augmented argument mapping. The AI can help an analyst with identifying hidden assumptions in their reasoning. [reference information to be provided when available]

Acknowledgements

The authors thank Simon Dennis, Ben Stone and Michael Diamond for their help with Study 2 as well as Andrew Robinson from CEBRA and Melanie Newfield from NZMPI for general assistance with the research. We thank the Intelligence and National Security reviewers for insightful and constructive feedback.

Disclosure statement

This paper describes evaluation of a training-based initiative. One author (van Gelder) delivered the training in a private capacity. The other authors (Kruger and Thorburn) were members of a research group directed by van Gelder.

Funding

This work was supported by the University of Melbourne Centre of Excellence for Biosecurity Risk Analysis (CEBRA) project #19NZ02.

Notes on contributors

Ariel Kruger is a research fellow at the Hunt Laboratory for Intelligence Research, University of Melbourne. His background is in the philosophy of science while his current research focuses on empirical investigations of intelligence analysis products and processes.

Luke Thorburn is a research associate specialising in statistics and machine learning at the Hunt Laboratory for Intelligence Research, University of Melbourne. He is also a PhD candidate at King's College London within the Centre for Doctoral Training in Safe and Trusted AI.

Timothy van Gelder is Director of the Hunt Laboratory for Intelligence Research in the Faculty of Science at the University of Melbourne. A philosopher by training, he is an applied epistemologist with experience in pure research, software development, consulting and training.

ORCID

Ariel Kruger http://orcid.org/0000-0002-0121-2780
Luke Thorburn http://orcid.org/0000-0003-4120-5056
Timothy van Gelder http://orcid.org/0000-0001-5606-495X

Data availability

The data that supports the findings of this study is openly available in the Open Science Repository at https://osf.io/g6r84/

Ethics

Study 2 was conducted with ethics approval from The University of Melbourne Faculty of Science Human Ethics Advisory Group ID: 2057037.1

Bibliography

Abrami, P. C., R. M. Bernard, E. Borokhovski, D. I. Waddington, C. A. Wade, and T. Persson. "Strategies for Teaching Students to Think Critically: A Meta-Analysis." *Review of Educational Research* 85, no. 2 (2015): 275–314. doi:10.3102/0034654314551063.

Alvarez, C. "Does Philosophy Improve Reasoning Skills?" Masters Thesis, University of Melbourne, 2007.

Barnett, A., T. Primoratz, R. de Rozario, M. Saletta, L. Thorburn, and T. J. van Gelder. *Analytic Rigour in Intelligence*. Melbourne, Australia: Hunt Laboratory for Intelligence Research, University of Melbourne, 2021. https://cpb-ap-se2.wpmucdn.com/blogs.unimelb.edu.au/dist/8/401/files/2021/04/Analytic-Rigour-in-Intelligence-Approved-for-Public-Release.pdf .

Battaglia, M. "Convenience Sampling." In *Encyclopedia of Survey Research Methods, by Paul Lavrakas. 2455 Teller Road, Thousand Oaks California 91320 United States of America*. Sage Publications, Inc, 2008. doi:10.4135/9781412963947.n105.

Biosecurity Act 1993 No 95, § 23. 1995. https://www.legislation.govt.nz/act/public/1993/0095/latest/DLM315284.html .

Chang, W., E. Berdini, D. R. Mandel, and P. E. Tetlock. "Restructuring Structured Analytic Techniques in Intelligence." *Intelligence and National Security* 33, no. 3 (2018): 337–356. doi:10.1080/02684527.2017.1400230.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic press, 1977.

Coulthart, S. "Why Do Analysts Use Structured Analytic Techniques? An in-Depth Study of an American Intelligence Agency." *Intelligence and National Security* 31, no. 7 November 9 (2016): 933–948. doi:10.1080/02684527.2016.1140327.

Coulthart, S. J. "An Evidence-Based Evaluation of 12 Core Structured Analytic Techniques." *International Journal of Intelligence and CounterIntelligence* 30, no. 2 April 3 (2017): 368–391. doi:10.1080/08850607.2016.1230706.

Cullen, S., J. Fan, E. van der Brugge, and A. Elga. "Improving Analytical Reasoning and Argument Understanding: A Quasi-Experimental Field Study of Argument Visualization." *Npj Science of Learning* 3, no. 1 December 4 (2018): 1–6. doi:10.1038/s41539-018-0038-5.

Davies, M., A. Barnett, and T. van Gelder. "Using Computer-Aided Argument Mapping to Teach Reasoning." In *Studies in Critical Thinking*. edited by A. Blair. Windsor Studies in Argumentation 8. Windsor, ON: WSIA, 2019. 131–176.

Davies, P. H. J., K. Gustafson, and I. Rigdin. "The Intelligence Cycle Is Dead, Long Live the Intelligence Cycle: Rethinking Intelligence Fundamentals for a New Intelligence Doctrine." In *Understanding the Intelligence Cycle*, edited by Phythian, M. 56–75. New York: Routledge, 2013.

Dhami, M. K., I. K. Belton, and D. R. Mandel. "The 'Analysis of Competing Hypotheses' in Intelligence Analysis." *Applied Cognitive Psychology* 33, no. 6 (November, 2019): 1080–1090. doi:10.1002/acp.3550.

Dorn, S. "Teaching Intelligence Analysis Writing Skills: A Program Evaluation." *Journal of Intelligence and Analysis* 24, no. 2 (2019): 73–94.

Dries, T., S. J. Handley, and M. F. Verde. "Fluency and Belief Bias in Deductive Reasoning: New Indices for Old Effects." *Frontiers in Psychology* 5 (2014): 631. doi:10.3389/fpsyg.2014.00631.

Dube, C., C. Rotello, and E. Heit. "Assessing the Belief Bias Effect with ROCs: It's a Response Bias Effect." *Psychological Review* 117 (2010): 831–863. doi:10.1037/a0019634.

Fechner, G. T., D. H. Howes, and E. Garrigues Boring. *Elements of Psychophysics*. Vol. 1. 1860. Reprint, New York: Holt, Rinehart and Winston, 1966.

Heuer, RJ, and RH Pherson. *Structured Techniques for Intelligence Analysis*. Cq Press, 2015.

Hulnick, A. S. "The Intelligence Producer – Policy Consumer Linkage: A Theoretical Approach." *Intelligence and National Security* 1, no. 2 May 1 (1986): 212–233. doi:10.1080/02684528608431850.

Immerman, RH. "Transforming Intelligence Analysis: 'The Tail that Wags the Dog." In *Rethinking Leadership and "Whole of Government" National Security Reform: Problems, Progress, and Prospects*, edited by Cerami , JR, and Engel, JA (Carlisle, Pennsylvania: Army War College (U.S.). Strategic Studies Institue), 2010. 73–110.

Immerman, R. H. "Transforming Analysis: The Intelligence Community's Best Kept Secret." *Intelligence and National Security* 26, no. 2–3 (April 1, 2011): 159–181. doi:10.1080/02684527.2011.559138.

Keren, G., and W. B. de Bruin. "On the Assessment of Decision Quality: Considerations regarding Utility, Conflict and Accountability." In *Thinking: Psychological Perspectives on Reasoning, Judgment and Decision Making*, edited by D. Hardman and L. Macchi, 347–363. Chichester, UK: John Wiley & Sons, Ltd, 2005. doi:10.1002/047001332X.ch16.

Kruger, A., T. van Gelder, and L. Thorburn. "The Impact of Evidence on Decision Making - Final Report." Center for Excellence in Biosecurity Risk Analysis, University of Melbourne, 2021.

Macmillan, Neil, A, and C Douglas Creelman. *Detection Theory: A User's Guide*. Mahwah, New Jersey: Lawence Erlbaum Associates, 2004.

MAF Biosecurity NZ. 2009. "Balance in Trade." https://www.mpi.govt.nz/dmsdocument/12576-Balance-in-trade .
MAF Biosecurity NZ. 2020. "Import Risk Analysis: Process Overview." https://www.mpi.govt.nz/dmsdocument/41779-Import-risk-analysis-process-overview .
Mandel, D. R. "The Occasional Maverick of Analytic Tradecraft." *Intelligence and National Security* 35, no. 3 (April 15, 2020): 438–443. doi:10.1080/02684527.2020.1723830.
Marrin, S. "Intelligence Analysis: Structured Methods or Intuition?" *American Intelligence Journal* 25, no. 1 (2007): 7–16.
Marrin, S. "Evaluating the Quality of Intelligence Analysis: By What (Mis) Measure?" *Intelligence and National Security* 27, no. 6 (December 1, 2012): 896–912. doi:10.1080/02684527.2012.699290.
Ministry for Primary Industries New Zealand. "Import Risk Analysis." Ministry for Primary Industries. Accessed March 9, 2021. https://www.mpi.govt.nz/import/importing-into-nz-how-it-works/import-health-standards/risk-analysis/ .
ODNI. 2015a. "Intelligence Community Directive (ICD) 203, Analytic Standards." Office of the Director of National Intelligence. https://www.dni.gov/files/documents/ICD/ICD%20203%20Analytic%20Standards.pdf .
ODNI *Rating Scale for Evaluating Analytic Tradecraft Standards with Amplified Guidance for Evaluators (Last Revised on 6 November 2015)*. Office of the Director of National Intelligence, 2015b.
Paolacci, G., J. Chandler, and P. G. Ipeirotis. "Running Experiments on Amazon Mechanical Turk." *Judgment and Decision Making* 5, no. 5 (2010): 411–419.
Phythian, M, and P. Gill. "From Intelligence Cycle to Web of Intelligence: Complexity and the Conceptualisation of Intelligence." In *Understanding the Intelligence Cycle*, edited by Phythian, M. 21–42. (New York: Routledge), 2013.
Pineda, P. "Evaluation of Training in Organisations: A Proposal for an Integrated Model." *Journal of European Industrial Training* 34, no. 7 August 31 (2010): 673–693. doi:10.1108/03090591011070789.
Poder, T. G., C. A. Bellemare, S. K. Bédard, J.-F. Fisette, and P. Dagenais. "Impact of Health Technology Assessment Reports on Hospital Decision Makers–10-Year Insight from a Hospital Unit in Sherbrooke, Canada: Impact of Health Technology Assessment on Hospital Decisions." *International Journal of Technology Assessment in Health Care* 34, no. 4 (2018): 393–399. doi:10.1017/S0266462318000405.
Professional Head of Intelligence Assessment. 2019. "Professional Development Framework: For All Source Intelligence Assessment." Crown Copyright.
Stanislaw, H., and N. Todorov. "Calculation of Signal Detection Theory Measures." *Behavior Research Methods, Instruments, & Computers* 31, no. 1 (1999): 137–149. doi:10.3758/BF03207704.
Stromer-Galley, J., P. Rossini, K. Kenski, B. McKernan, B. Clegg, J. Folkestad, C. Østerlund, et al. "Flexible versus Structured Support for Reasoning: Enhancing Analytical Reasoning through a Flexible Analytic Technique." *Intelligence and National Security* 36, no. 2 (February 23, 2021): 279–298. doi:10.1080/02684527.2020.1841466.
Thomason, N. "A012. Critical Thinking Final Report." University of Melbourne, 2014.
Toulmin, S. E. *The Uses of Argument*. 2nd ed. Cambridge: Cambridge University Press, 2003.
Van Gelder, T. J. "The Rationale for Rationale™." *Law, Probability and Risk* 6, no. 1–4 (2007): 23–42.
Van Gelder, T. J. "Argument Mapping." In *Encyclopedia of the Mind*, edited by H. Pashler. Thousand Oaks, CA: SAGE. 50–52, 2013.
Van Gelder, T. J. *The Reasoning PowerPoint App*. Melbourne, Australia: Austhink Consulting Pty Ltd, 2016. http://learn.vangeldermonk.com/courses/reasoning-app .
Van Gelder, T. J. "SWARM Project Final Report." February 2020.
Van Gelder, T. J., M. Bissett, and G. Cumming. "Cultivating Expertise in Informal Reasoning." *Canadian Journal of Experimental Psychology* 58, no. 2 (2004): 142–152. doi:10.1037/h0085794.
Walsh, P. F. 2017. "Teaching Intelligence in the Twenty-First Century: Towards an Evidence-Based Approach for Curriculum Design." *Intelligence and National Security*, June 30, 1–17.
Walton, D., C. Reed, and F. Macagno. *Argumentation Schemes*. Cambridge: Cambridge University Press, 2008. doi:10.1017/CBO9780511802034.
Warren, N. L., M. Farmer, G. Tianyu, and C. Warren. "Marketing Ideas: How to Write Research Articles that Readers Understand and Cite." *Journal of Marketing* 85, no. 5 September 1 (2021): 42–57. doi:10.1177/00222429211003560.
Whitesmith, M. "The Efficacy of ACH in Mitigating Serial Position Effects and Confirmation Bias in an Intelligence Analysis Scenario." *Intelligence and National Security* 34, no. 2 February 23 (2019): 225–242. doi:10.1080/02684527.2018.1534640.
Zieliński, W. 2019. "A New Exact Confidence Interval for the Difference of Two Binomial Proportions." *ArXiv Preprint ArXiv:1903.03327*.