THE IMPACT OF EVIDENCE ON DECISION MAKING - FINAL REPORT

CEBRA Project 19NZ02 Deliverable 3

15 August 2021

Dr Ariel Kruger

A/Prof. Tim van Gelder

Luke Thorburn

The Hunt Laboratory for Intelligence Research, The University of Melbourne

Executive summary

BACKGROUND

In 2015, the New Zealand Ministry of Primary Industries (MPI) commenced a series of activities aimed at improving the clarity and rigour of reasoning in Import Risk Assessments (IRAs). These activities consisted in providing training for staff, mostly in the Plants division, in a generic method for articulating and presenting complex reasoning known as CASE ("Contention, Argument, Evidence, Source"). Other activities included group practice sessions, and encouraging staff to apply the method when drafting IRAs. We call this series of activities the **CASE initiative**.

In 2019, MPI decided to evaluate the CASE initiative, engaging CEBRA to undertake the current project. The focus was to be on the impact of CASE on MPI decision making, aiming specifically to "identify and characterize the impact, on Import Health Standards (IHS) decisions, of recent changes in the presentation of evidence and arguments," and to make recommendations regarding the continuation or modification of the initiative.

At the outset of this project it was not clear how the impact on decision making of something like the CASE initiative could be assessed. Preliminary literature searches suggested that the project would need to be methodologically innovative. This, and other complications meant that the project ended up being exploratory in nature.

OUR RESEARCH

After an extensive literature review, we proceeded by dividing the topic into two main research questions: (1) to what extent has the CASE initiative improved the clarity and rigour of reasoning in IRAs? and (2) to what extent has this improvement led to better decisions?

To address the first question, we conducted two studies.

Study 1 aimed to assess the extent to which IRAs produced after the start of the CASE initiative (which we call "post-CASE") showed stronger CASE structure than those produced previously ("pre-CASE"). We took a purposeful sample of two pre- and two post-CASE IRAs, and randomly sampled sections of reasoning from them. We coded these sections of reasoning for the presence of CASE structure. We found much stronger CASE structure in the post-CASE samples, especially in one IRA (Prunus). Taking into account methodological concerns, we cautiously infer to greater CASE structure in post-CASE IRAs in Plants division more generally.

Study 2 sought independent confirmation that post-CASE IRAs were better reasoned. It aimed to assess the extent to which post-CASE IRAs would be judged as better-reasoned, not by experts, but by general readers. For our sample of general readers we recruited workers on Amazon Mechanical Turk. These participants were presented with many pairs of sections of reasoning, one from a pre-CASE IRA and one from a post-CASE IRA, and asked to make a "forced choice" of the one that was better reasoned and communicated. We found that these general readers showed a small but statistically robust preference for post-CASE reasoning. Taking into methodological concerns, we regard this as providing weak evidence for a small improvement in the eyes of general readers.

The two studies provide convergent evidence that post-CASE Plants IRAs have greater clarity and rigour. While our studies were not well-suited to isolating causal factors, on qualitative grounds we regard this difference as plausibly being largely if not wholly attributable to the CASE initiative.

To address the second question – to what extent has this improvement led to better IHS decisions? – we conducted one study.

Study 3 did not attempt to assess improvement in the overall quality of IHS decisions, which for both practical and theoretical reasons could not be assessed. Rather, it aimed to assess improvement in one critical aspect of decision quality, the *alignment* of IHS decisions with their corresponding risk assessments. We coded a selection of decisions from post-CASE IHS reports for the severity of the measures they required; coded the corresponding risk assessments for the level of risk; and coded each decision-assessment pair for the match between severity and risk. We found that alignment was high in all IHS reports in our sample. Post-CASE decisions were more aligned, but this was not statistically significant.

Due to the marginal difference in alignment in our sample, methodological concerns, and alignment being only aspect of good IHS decisions, we cannot infer that post-CASE IHS decisions were better.

OUR CONCLUSIONS

Overall, our research indicates that the CASE initiative improved clarity and rigour in Plants IRAs, but is inconclusive with respect to its impact on the quality of IHS decisions.

One important finding from Study 1 is that the Prunus IRA exhibited CASE structure almost perfectly. This demonstrates that MPI *can* produce IRAs which, among other virtues, very consistently adhere to certain fundamental principles of good reasoning and communication.

A noteworthy finding from Study 3 was that IHS decisions are well-aligned with the corresponding risk assessments. This may be as expected, but our research confirms and quantifies this fact.

At a methodological level, we believe that the research designs used in Studies 1 and 2 are essentially fit for purpose. With modifications to overcome some limitations in our exploratory studies, they could be used in larger confirmatory studies. The design of Study 3 is less promising. We briefly describe some alternative research approaches in the final section.

OUR RECOMMENDATIONS

We make four recommendations.

1. Continue the CASE initiative. The CASE initiative (suitably strengthened) should be continued, because: the original reasons for introducing it still stand; our research finds a positive impact on IRA clarity and rigour, and does not rule out impact on decision making; the initiative is likely to have benefits other than improved decision making; and we are not aware of any other strategy which might be more effective at reasonable cost.

If continuing the initiative:

2. Strengthen the implementation. A strengthened CASE initiative would likely have stronger impact. It could be strengthened by making CASE activities more regular; developing a handbook; verifying that staff are achieving suitable proficiency; and including CASE adherence in the IRA review process.

3. Tailor the CASE approach. The generic CASE approach could be tailored to better suit MPI's context and needs, through steps such as developing templates for specific uses such as IRA drafting.

4. Integrate evaluation. Some forms of evaluation should be integrated with the initiative itself, rather than being done retrospectively.

Acknowledgments

The authors are grateful to Melanie Newfield for help in collecting and collating the data we needed, finding required information when requested and providing insight into the decision-making processes at NZMPI. We are also grateful to Professor Simon Dennis and his team: Dr Ben Stone and Michael Diamond for their substantial contribution to Study 2: Do general readers find post-CASE IRAs better-reasoned? Professor Andrew Robinson provided important guidance in research design and statistical analysis.

Table of Contents

Tables

Figures

Acronyms and technical terms

Term	Definition
CA	Component Argument – A subsection of an Import Risk Analysis report presenting reasoning in support of a judgement, such as a judgement of likelihood of entry
CASE	"Contention, Argument, Evidence, Source" – The core elements of an approach to structured argumentation
CEBRA	Centre of Excellence for Biosecurity Risk Analysis, University of Melbourne
DAWE, DAWR	Australian Department of Agriculture, Water and the Environment
EIDM	Evidence-informed decision making
HTA	Health Technology Assessment
IHS	Import Health Standard
IRA	Import Risk Assessment
MPI	New Zealand Ministry for Primary Industries
RMP	Risk Management Proposal
PEQ	Post-entry quarantine.
Pre-CASE	Before the start of the CASE initiative at MPI in 2015
Post-CASE	After the start of the CASE initiative in 2015

1 Introduction

One challenge for organisations charged with biosecurity management is assessing the risks associated with importing products into the country. New Zealand's Ministry for Primary Industries handles this, in part, by developing substantial documents known as Import Risk Analyses (IRAs). An IRA considers a kind of import (e.g., fresh fruit/vegetables) and the various hazardous pests or diseases associated with that kind of import. It provides an assessment of the risk posed by each hazard, supported by evidence and arguments.

IRAs are important documents. They help inform the difficult policy decisions around risk mitigation, which aim to balance minimizing the potential harms from introduced pests and diseases with other objectives such as promoting trade. They are also circulated to stakeholders such as importers, helping explain and justify MPI's assessments and policy decisions.

For various reasons, it is essential that the reasoning in IRAs be both rigorous and easily understood. Good reasoning should help ensure that the risk assessments are sound; it should make them more credible to decision makers, and thus more likely to be weighted appropriately; and it should lead to greater acceptance of the assessments by stakeholders.

MPI uses a number of strategies to ensure that IRAs are suitably high quality, including providing training for risk analysts, and having IRAs undergo review processes. In 2015, the organisation started supplementing those strategies with another kind of training, focusing on general principles for articulating and communicating reasoning. It began offering analysts one-day workshops in argument mapping, using an approach known as "CASE,"[1] provided by an outside facilitator. In addition, groups of staff got together to work on their skills, and some managers encouraged their teams to draft their reasoning in accordance with CASE principles.

This effort, which we call the "CASE initiative," evolved over subsequent years in a largely organic and *ad hoc* manner. By 2019, it seemed clear that the approach had some enthusiastic internal supporters, and was having at least some impact on some IRAs. However the initiative was also imposing some cost on the organisation. Was the investment warranted? This was unclear, since the CASE initiative had not been undergoing any systematic evaluation.[2]

Thus, in 2019 MPI approached CEBRA for assistance, leading to the project described in this report. The evaluation was to look at outcomes for the organisation, rather than at the professional development of, or benefits, for, the individuals involved. The focus was to be on the relationship between the presentation of reasoning in reports, and the decision making informed by those reports.[3] Specifically, was the CASE initiative helping improve decision making at MPI? And should the initiative be continued, or modified?

---

[1] "CASE" is an acronym for "Contention, Argument, Evidence, Source." Argument mapping, and CASE, are described in the Background section.

[2] Organisations very often evaluate training by having participants fill out post-session evaluation forms. They rarely do anything more in-depth or rigorous, due in part to difficulty and cost of conducting such evaluation (Pineda, 2010). To our knowledge, MPI was not making use of post-session participant evaluation.

[3] The original project brief specified that "The research approach will be an exploratory, retrospective, qualitative analysis aiming to identify and characterize the impact, on Import Health Standards decisions, of recent changes in the presentation of evidence and arguments."

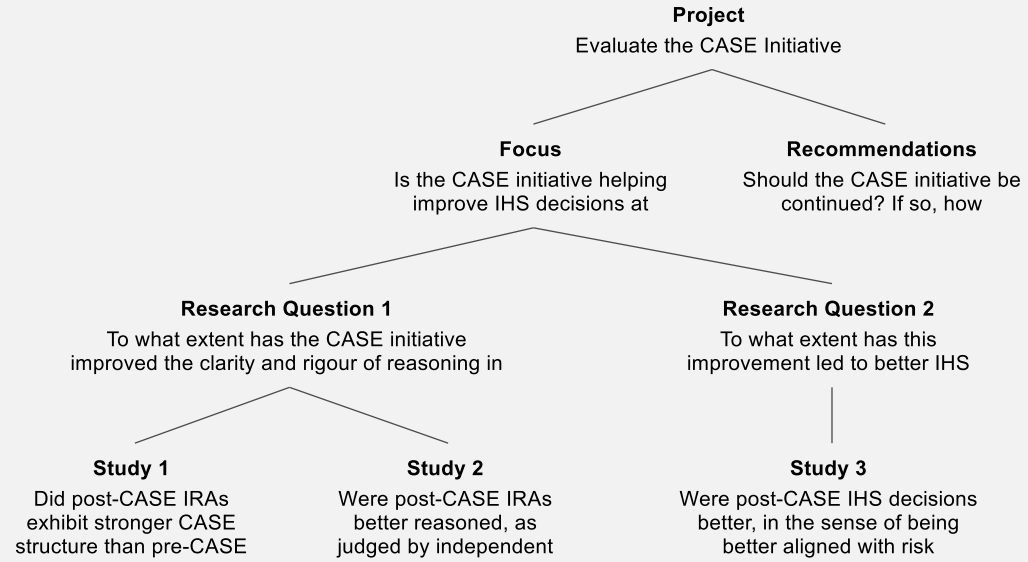
It sounds reasonable to ask whether the CASE initiative was helping improve decision making, but the question conceals a profound difficulty. It supposes we can know the quality of MPI decisions before, and after, the start of the CASE initiative. The problem is that there is no straightforward way of measuring the overall quality of complex, real-world decisions. Indeed, there is not even any agreement about what quality *is* for such decisions. Given this problem, it is not surprising that initial searches turned up no evidence of any prior research, in any field, on how the presentation of reasoning in reports affects the quality of decisions informed by those reports.

The project, it seemed, would need to break new ground. It was therefore conceived as exploratory in nature. It would develop and test an approach to tackling the core question, with two expected outcomes: better understanding of how to tackle such problems, and at least initial findings, which might serve as hypotheses to be tested in subsequent, more ambitious research, and meanwhile support at least provisional guidance for MPI with regard to the CASE initiative.

The project commenced in early 2020.[4] As expected in an exploratory exercise, our understanding of the problem, and methods we might use, evolved over the course of the project. We ended up breaking the overarching objective, to assess whether the CASE initiative is helping improve decision making, into two main research questions:

1. To what extent has the CASE initiative improved the clarity and rigour of reasoning in IRAs?
2. To what extent has this improvement (if any) led to better decisions?

Figure 1-1: Overall structure of our research project, showing the breakdown of questions and studies.



[4] The project got underway just as COVID was becoming a pandemic, and states were introducing lockdowns and border controls. This precluded anticipated travel by the researchers in Melbourne to MPI in Wellington, which in turn had some impact on project design. For example it ruled out in-person meetings with MPI staff including decision makers.

To help answer these questions, we conducted a literature review, and three pilot studies, addressing three more specific questions:

1. Did post-CASE IRAs exhibit stronger CASE structure than pre-CASE IRAs?
2. Were post-CASE IRAs better reasoned, as judged by independent readers?
3. Were post-CASE policy decisions better, in the sense of being better aligned (explained below) with the risk assessments in the relevant IRAs?

The first two studies addressed the first major question; the third study addressed the second question. The relationships among these questions and studies are shown in Figure 1-1.

Our studies took place within the Plants division of MPI, where the CASE initiative was mainly located.

Study 1 aimed to ascertain whether post-CASE IRAs had stronger CASE structure. We took samples of IRAs from before, and after, the start of the CASE initiative; and from these IRAs, we drew systematic samples of the chunks of reasoning within them. We developed a qualitative coding scheme for measuring the extent to which each chunk exhibited CASE structure, and applied it to obtain a CASE score for each chunk. We then compared the pre- and post-CASE scores using simple descriptive statistics, and by developing a mixed-effects model.

We found that the post-CASE IRAs in our sample did in fact have much stronger CASE structure. This suggested that the CASE initiative was working, at least in the sense of changing the way chunks of reasoning were being drafted. There are theoretical reasons for thinking that this change amounted to better reasoning, but it is a further question whether readers, such as policy makers or stakeholders, would recognize this difference as improvement.

Therefore, in Study 2, we took the issue to an independent jury. We recruited study participants on a commercial cloud platform. Each participant was shown many pairs of chunks of reasoning. Each pair contained one pre-CASE chunk and one post-CASE chunk, without any indication of which was which. Participants were asked to select the chunk which was “better reasoned and communicated,” and to provide a brief written comment supporting their choice. As in Study 1, we analysed the resulting data with descriptive statistics and with mixed effects modelling.

We found that these independent judges, assumed to have no special knowledge of biosecurity, or CASE, tended to regard the post-CASE chunks as better. This coheres with the finding from Study 1. Jointly, the studies suggest that the post-CASE IRAs in our sample really did have somewhat greater clarity and rigour.

That set the stage for addressing the primary question in this project, whether the CASE initiative was helping improve policy decisions. To make this tractable, we focused on a particular aspect of decision quality, which we called *alignment*. This is the extent to which the severity of the risk management measures specified in policy decisions match the levels of risk in the corresponding IRAs. Other things being equal, there should be alignment, which is just a way of saying that risk management measures should be proportionate to the risks involved.

To assess whether post-CASE alignment was better, we took samples of pairs of reports, from before and after the CASE initiative commenced. Each pair contained an Import Health Standards document, which presents the decisions on risk management measures, and its corresponding IRA. We coded the IRAs for the level of risk associated with each hazard, and coded the IHSs for the severity of the decisions regarding risk management measures. We then classified each decision as being aligned, or not aligned, with the level of risk. We analysed the results using a binomial regression model. We

found that post-CASE decisions were, in fact, better aligned with the corresponding risk assessments, though this difference was statistically marginal.

These results are broadly positive. In all three studies, the data were consistent with the CASE initiative having improved reasoning in IRAs, and improving policy decisions. However we are not in a position to firmly assert that the CASE initiative had these effects. Our studies all had limitations in methodology, and in obtaining sufficient data. We therefore reiterate the exploratory nature of the project. Our three studies are best regarded as pilots, yielding preliminary findings, and considerable insight into how more rigorous and extensive studies could be conducted.

The remainder of this report is structured as follows.

- In Section 2, Background, we go into more depth into topics the biosecurity decision making process at MPI, and the CASE approach and its implementation at MPI.
- Section 3 presents findings from our review of previous literature.
- Sections 4, 5 and 6 each describe one of our three pilot studies.
- Section 7, Conclusion, draws together our findings across the three studies; makes some recommendations regarding the CASE initiative; and indicates some directions for future research.

Appendices provide detailed information on various topics, removed from the main body to improve readability.

Ethics

For our first study (Study 1: Do post-CASE IRAs show stronger CASE structure?) and our third study (Study 3: Are post-CASE IHS decisions better aligned with risk assessments?) ethics approval was not required as the research did not involve human subjects.

Our second study (Study 2: Do general readers find post-CASE IRAs better-reasoned?) involved human subjects and we obtained University of Melbourne ethics approval (ID: 2057037.1).

Data availability

All data is publicly available at https://osf.io/g6r84/. No sensitive information is contained in this repository.

2 Background

This section covers the background to our project and the three studies, including:

- How risk assessments inform risk management decisions at MPI;
- The initiative to improve IRAs;
- The CASE method; and
- How MPI introduced CASE.

2.1 Informing biosecurity risk decisions at MPI

2.1.1 Overview

MPI's risk management decision making process starts with external stakeholder or MPI managers identifying the need for a new or amended decision document (IHS) regarding the importation of a commodity. MPI then begins drafting an Import Risk Analysis (IRA), which analyses the risks posed by pests associated with a commodity. Decision makers at MPI use the IRA to inform their decisions on what measures are appropriate to manage those risks. Risk management decisions are then outlined in an Import Health Standard (IHS), which is a legal document that specifies what risk management measures are required when importing a particular commodity.

2.1.2 The Import Risk Analysis (IRA)

An IRA "assesses the pest and disease risks associated with importing a wide range of plants, animals, and other products" (Ministry for Primary Industries New Zealand, 2020b). Its development is a crucial part of process of making risk management decisions, as it helps to inform what management measures are appropriate. During its development the IRA is reviewed and modified by external stakeholders as well as internal MPI risk managers.

The report is initiated by the Chief Technical Officer (CTO) with the aim of informing risk management requirements for importing a commodity. The report development process begins with the compiling of a list of potentially hazardous pests and diseases using databases and libraries. Each pest or disease on the list is then assessed against criteria to determine if it constitutes a 'hazard' warranting further assessment.

That further assessment is an analysis of risk. Legislation requires that a risk analysis considers following factors (Biosecurity Act 1993 No 95, 1995):

1. The likelihood that importation of the commodity will also import organisms:
2. The nature of the organisms that may be imported:
3. The possible effect on human health, the New Zealand environment, and the New Zealand economy of those organisms:
4. New Zealand's obligations under international agreements other than the SPS agreement.

Thus, at a minimum, the IRA must evaluate risk along these dimensions. In practice, IRAs consider these factors and more besides. What else is considered can depend on the commodity being imported. For example, in nursery stock IRAs, the likelihood that a pest or disease will be detected in post-entry quarantine is also considered.

For some IRA, measures that might be capable of managing risk are also assessed with the aim of evaluating the level of risk reduction achieved by those measures. They are then drafted as proposals for consideration by decision makers.

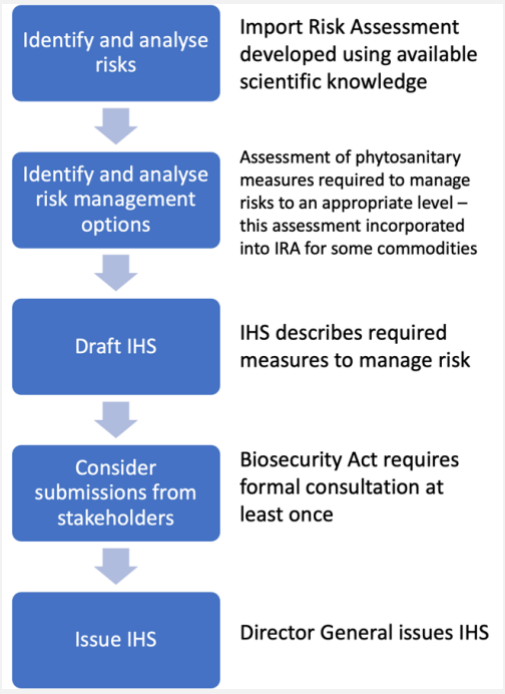
Once the CTO is satisfied the that IRA has met its legislative and other requirements, it can then be used to inform the development of a Risk Management Proposal (RMP) and an IHS.

2.1.3 The Import Health Standard (IHS)

An IHS is a legal document detailing the "biosecurity requirements that commodities must meet before biosecurity clearance can be given and the consignment can be successfully imported into New Zealand" (Ministry for Primary Industries New Zealand, 2020a). In other words, it details the *decisions* that have been reached regarding risk management measures that are applied to an imported commodity. According to S.23 of the Biosecurity Act 1993, decisions on what requirements are appropriate must have regard to the risks associated with the import as well as matters not directly related to biosecurity. The other matters include: the direct cost of the requirements on imports, the direct cost of the requirements on the Crown and technical and operational factors involved in implementing the requirements (Biosecurity Act 1993 No 95, 1995)[5].

Requirements can be separated into two categories, those that must be met prior to the import arriving in NZ and those that are met subsequent to it. The type of requirements imposed prior to arrival in NZ depend on the type of commodity. For example, fresh fruit/vegetables requirements include inspection, inert packaging, testing for regulated pests and specific phytosanitary measures. While requirements for nursery stock imports must either use integrated measures or come from a production are free from all regulated pests.

Figure 2-1: Flowchart of the decision-making process at MPI (Simplified from a simplified the flowchart available on the Biosecurity NZ website (IHS Development Process, n.d.)



Once the commodity arrives in NZ, another set of requirements must be met and these again depend on the commodity being imported. For example, if importing fresh fruit/vegetables, then the requirements that are imposed are actions to be taken should a pest be intercepted on a consignment at the border. These actions vary in severity depending on the level of risk posed by the pest. On the other hand, if importing nursery stock, then the requirements imposed are diagnostic tests that must be conducted on the stock during a mandatory quarantine period. All stock is visually inspected during quarantine, but pests that pose a significant risk or are unlikely to be spotted, will require more sensitive and specific testing than visual inspection alone.

Once the CTO believes that a set of requirements could effectively manage the risks, and satisfies other requirements detailed in the Biosecurity Act, they may then draft a proposed IHS. The CTO is mandated to consult "every department whose responsibilities for human health or natural resources may be adversely affected by it; and any other person the officer considers to be

[5] For a full list of considerations, see 9.2: Full list of legislated considerations

representative of the classes of persons having and interest in it" (Biosecurity Act 1993 No 95, 1995). After consultation and review, the Director General of MPI can then issue the IHS.

2.2 Improving Import Risk Analyses

Around 2014-15, there was interest in improving the quality of import risk analyses in both the New Zealand Ministry of Primary Industries and the (then) Australian Department of Agriculture and Water Resources. Our understanding is that the ambitions of both organisations included:

1. Improving the quality of risk assessment *judgements*;
2. Improving the clarity and rigour of the presentation, in IRAs, of the reasoning behind those judgements; and
3. Increasing the efficiency of production of IRAs and documents.

The second of these, improving the clarity and rigour of the presentation of reasoning, was thought important for at least three reasons:

1. It would likely contribute to the quality of judgements;
2. It would likely help improve decision making, since the judgements would be more credible, and the reasoning more persuasive, to decision makers;
3. It would help communicate the thinking to external stakeholders.

After discussions with CEBRA, both organisations decided to try improving the presentation of reasoning in IRAs by providing training to some staff in the CASE approach.

2.3 The CASE approach

2.3.1 What is CASE?

The acronym CASE stands for "Contention, Argument, Evidence, Source."[6] In essence it is an *argument scheme*, a template representing a common pattern of reasoning or argumentation (Walton, Reed, & Macagno, 2008). However CASE also refers to a more complex, dynamic template and a larger body of theory and techniques for articulating, structuring, strengthening and presenting reasoning on almost any topic.

2.3.1.1 The CASE scheme

For example, here is a small sample of real-world reasoning:

> *The Reserve Bank is unlikely to lower interest rates. According to the most recent Treasury report, economic growth is currently almost 3% per annum. This is comfortably above the RBA's growth target.*

[6] In the acronym, Evidence and Source are switched to get CASE rather than CAES. The two are pronounced the same way.

This reasoning has the CASE elements, as shown in Table 2-1:

Table 2-1: Elements of the core CASE scheme, and their definitions, as applied to a simple example.

Contention	The Reserve Bank is unlikely to lower interest rates.	The main point the reasoning establishes - or at least purports to establish.
Argument	Economic growth is comfortably above the RBA's target growth rate.	A major point being put forward as either supporting the contention (a **Reason**) or opposing it (an **Objection**).
Evidence	Economic growth is currently almost 3% per annum.	A specific piece of information backing up the Argument. **Counter-evidence** is information opposing the Argument.
Source	The most recent Treasury report	The source of the information.

Note that the CASE components in the original text are "jumbled up" relative to ordering in the CASE scheme. This is very common. The diversity of orderings of logical components in ordinary prose helps explain why reasoning is often hard to follow.

2.3.1.2 More complex CASE structures

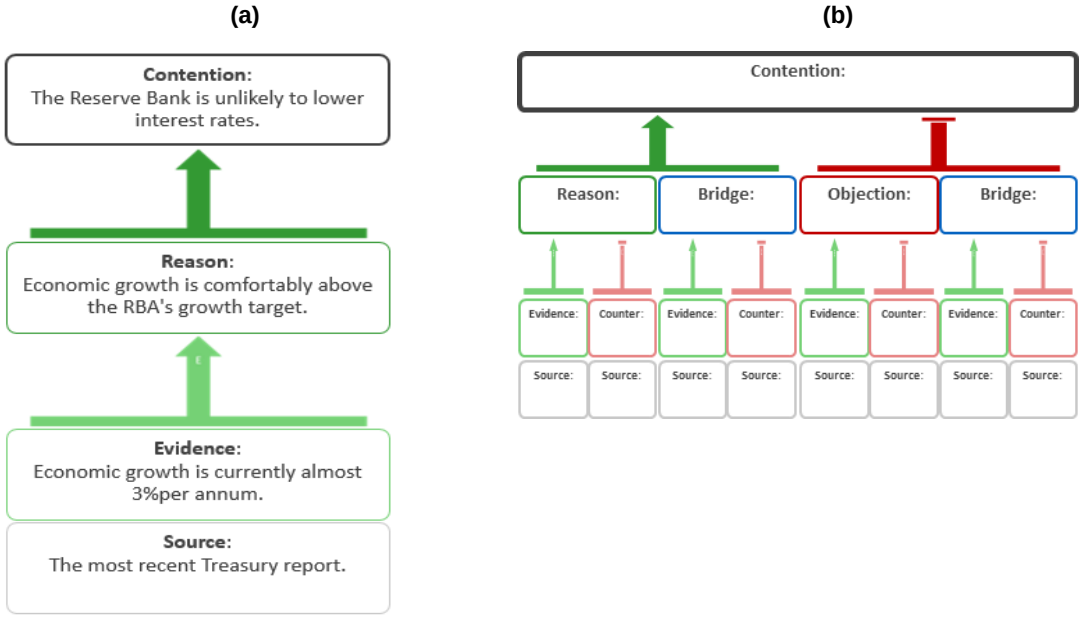
The CASE scheme shown above is just the core of a more complex, dynamic framework for structuring reasoning. Additional elements include:

- Conflicting or opposing Arguments (Objections, as opposed to Reasons) and Evidence (Counter-Evidence)
- More than one Argument bearing on a Contention, and more than one Evidence item bearing on an Argument
- Multiple levels of Argument between Evidence and Contention
- The use of *abstraction* to add order and strength;
- The compound nature of reasoning units, where both Arguments and Evidence items require additional claims called Bridging claims. When not made explicit, these Bridging claims are hidden assumptions.

2.3.1.3 CASE and argument mapping

The CASE scheme is often combined with argument mapping (also known as argument diagramming or argument visualisation). Figure 2-2 (a) shows a mapped version of the reasoning in the example above. The benefits obtained from representing the CASE structure diagrammatically are modest in this example, but increase as CASE-based reasoning gets more complex. Figure 2-2 (b) shows in schematic form many of the ways in which CASE elements can combine into larger structures.

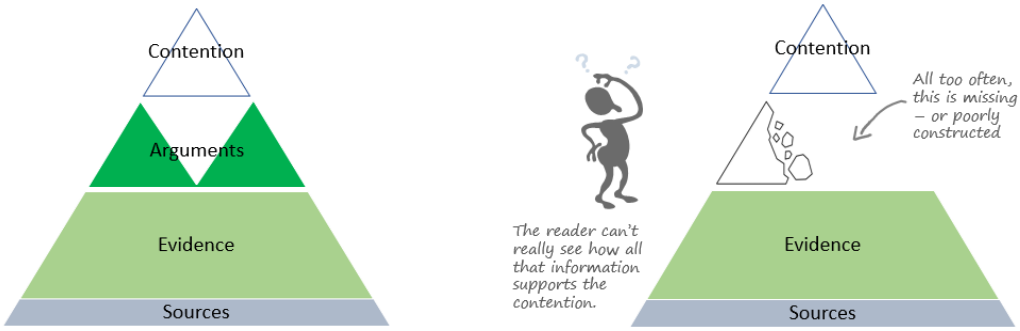
Figure 2-2: Examples of the CASE scheme in argument mapping format. (a) A simple piece of reasoning illustrating the core CASE scheme. (b) An expanded version of the CASE scheme, showing the major kinds of relationships between reasoning units. These diagrams are produced using the Reasoning add-in for Microsoft PowerPoint developed specifically for use in the DAWR and MPI CASE training.



2.3.1.4 Missing or poorly-developed "Argument" level

One of the skills emphasized in CASE training is properly articulating the Argument level – i.e. the Reasons or Objections most generally and directly supporting or opposing the main Contention. One reason for this emphasis is that failure to articulate the Argument level properly is a very common problem. The Arguments are either missing or inadequately expressed, as illustrated in Figure 2-3.

Figure 2-3: Illustration of the failure to properly articulate the Argument level. Graphics drawn from the CASE training materials provided to MPI training participants.



The problem manifests in texts that look like an information dump with a conclusion attached (easily caricatured as "blah blah blah… therefore X"). It can be very difficult for a reader to see exactly how the information is supposed to support that conclusion.

2.3.1.5 CASE origins

CASE was put together by one of the authors of this report, Tim van Gelder, over a period of many years in which he was engaged in both teaching reasoning and critical thinking to undergraduates,

and running workshops (both training, and facilitation) with organisations to improve on-the-job reasoning and writing.[7] He was drawing on prior work of various kinds, including.

1. The considerable work over many decades of researchers and teachers in the field of informal logic. This work included early methods for diagramming reasoning, e.g., (Scriven, 1976).
2. The development of argument mapping methods, and software to support argument mapping (Kirschner, Buckingham Shum, & Carr, 2002; Okada, Buckingham Shum, & Sherborne, 2008). Work at the University of Melbourne in the late 1990s and early 2000s using argument mapping to support the development of critical thinking skills was an important part of this effort (Rider & Thomason, 2008; van Gelder, 2002; van Gelder, Bissett, & Cumming, 2004).
3. The "Pyramid Principle" developed by corporate trainer Barbara Minto, and presented in her book of the same name (Minto, 2009). This book is widely used for structuring and presenting reasoning in the management consulting industry. The Pyramid Principle's emphasis on abstraction was incorporated into CASE.[8] It is discussed further in Section 3.6.

2.3.1.6 Scope and limits of CASE

CASE constitutes some of the most basic knowledge that any knowledge worker should have about the structure of reasoning generally, and how to make that structure explicit. However CASE is far from the whole story about good reasoning. It is focused on the fundamentals of argument structure. It does include some guidelines for strengthening reasoning. However, reasoning with well-articulated CASE structure might still be seriously flawed. Skilled practitioners draw on a much larger body of theory, practical guidelines, and know-how to eliminate or reduce errors.

2.3.2 Using CASE to develop and present reasoning

The standard way to use CASE to develop and present reasoning involves three main steps.

First, the reasoning is developed in argument mapping format. Following CASE guidelines, the reasoning is articulated, refined and strengthened.

Second, the reasoning is transferred into CASE-structured prose. Here is the simple example used above, in strict CASE-structured prose format:

> [Contention] The Reserve Bank is unlikely to lower interest rates.
>
> [Argument] Economic growth is comfortably above the RBA's target growth rate.
>
> - Economic growth is currently almost 3% per annum. (The most recent Treasury report.)

Note some key features:

- The main Contention is at the top.
- The Argument is nested under the Contention.

---

[7] CASE has not as yet been described in any substantial publication. Very brief descriptions are found in (van Gelder, 2016, 2019). Fuller presentations are found in the materials used in DAWR training workshops, including an online short course (http://learn.vangeldermonk.com/courses/argument-mapping), the workshop exercise booklet, and the workshop slides. Access to these might be obtained by contacting the authors.

[8] For more on the Pyramid Principle, see Section 3.6

- Items of Evidence are bulleted, and are nested directly underneath the Argument on which they bear.

Advantages of this mode of presentation include:

- It is always easy and fast to find the main point being argued for;
- It is easy to see what Arguments have been raised;
- It is easy to see what items of information have been identified as Evidence, and which Arguments they support. Conversely, for any Argument, it is easy to see what Evidence is being provided to back it up.

Third, the "raw" CASE-structured prose is further massaged to make it more reader-friendly, while retaining the benefits of the CASE structuring.

For a complex example of applying CASE to improve the structure and presentation of reasoning in the context of real biosecurity risk analysis, see Appendix 4.

CASE users need not always follow the three steps just outlined. A skilled practitioner can go more or less directly to step 3, drafting readable prose with strong CASE structure.

2.4 Introducing CASE in MPI

CASE was introduced to MPI by two main methods: one-day intensive training workshops led by a CASE expert, and additional practice sessions, led by a MPI manager who was an internal champion of the approach, and proficient in CASE.[9]

Seven one-day training workshops were held, on the following dates:

- 17 & 18th September 2015
- 28 February & 1 March 2016
- 12, 13 & 14th December 2018

The workshops were conducted in in-person sessions in Wellington. Most workshops were a standard one-day introduction, but some focused on deepening expertise.

Each workshop had around 10-15 participants, but a portion of participants attended multiple workshops, so we estimate around 50-60 staff received introductory training. Most attendees were analysts or managers involved in the development of Plants division IRAs.

Workshops covered the essentials of CASE, as outlined in Section 2.3.1, and involved many practice exercises. Most examples and exercises were thematically related to biosecurity risk, and many exercises were directly related to IRAs.

A similar series of training workshops were being held at DAWR over the same period. The training was iteratively improved over the 4-5 year period, based on experience running workshops in both contexts. Improvements included finding better biosecurity-related examples to use in exercises, developing better exercise activities, such providing greater scaffolding to help guide answer development, and developing engaging quizzes using real-time audience response technology.

The steps described above constitute a serious effort to develop staff skills and to improve the quality of reasoning in IRAs. That said, the initiative could have been more extensive and rigorous.

[9] The manager's proficiency was developed through their extensive involvement in the initiative, including organising and participating in many of the workshops, organising and leading practice sessions, leading the introduction of CASE in report drafting, and mentoring other staff.

Not all staff were exposed to the training and practice, and not all of those who were exposed were able to master CASE, given the (necessarily) limited time and resources invested. Although the CASE theory is basic, applying it in cases of moderate, or greater complexity can be quite challenging. As with any serious skills, mastery requires considerable effortful practice with good feedback (Ericsson & Pool, 2016). A one-day training workshop is not enough on its own; it can only instil understanding and proficiency in the more elementary aspects of CASE. It was clear to the instructor, in these workshops as in similar workshops in many other contexts, that some participants had greater prior aptitude than others, and that the level of understanding and skill possessed at the end of the training sessions varied widely. To achieve widespread proficiency in the full CASE method (or any other approach to improving logical reasoning and written presentation) would require a more intensive training and development program than was undertaken, and may need some sort of certification process. These issues are discussed further in Section 7.2, Recommendations.

3 Literature Review

3.1 Introduction

This preview provides context for our studies by synthesizing literature relevant to the questions listed below. We used a narrative approach ("Narrative Literature Review," 2017), identifying literature via database searches and following citation chains. The review scope is defined by the questions and not limited to a particular domain such as biosecurity.

1. What are the common barriers to proper use of evidence and arguments to inform decisions? Could those barriers be overcome with something like CASE?
2. What interventions have been used to improve how evidence/argument is used to inform decisions? Are any similar to CASE? How effective are they?
3. What methods already exist to measure the impact evidence/argument has on decision making? Are any similar to our methodology, or potentially usable in our project?
4. To what degree is structured argumentation used in reports generally? What benefits can be derived from this use?

The answer to question 1 will help frame the context in which the CASE Structure could be applied. When properly applied, CASE guidelines can strengthen arguments which entails using evidence in support of the argument. Thus we sought to survey the literature to see if, and why, arguments and evidence are neglected in a variety of contexts. We might then learn about specific motivations to use the CASE Structure.

Question 2 sought to determine what interventions already exist to improve the use of evidence and argument. If those interventions are designed to ameliorate the same kind of difficulties as CASE and are successful at it, then we can show how CASE fits in to that broader literature on interventions.

Question 3 specifically address methods to measure the impact that evidence and argument have on decision making. To know if it makes a difference, and how much of a difference, you have to be able to measure it. This part of our literature view tried to locate existing methods which make that measurement as it was also a crucial part of our study. If we found methods purporting to measure that difference then they could be used to inform our own methodology.

The final question provides context for the CASE approach. A plethora of structured argumentation techniques have been developed in various disciplines and it is important to see how CASE sits in that context. Moreover, CASE itself is derived from earlier theory on structured argumentation, which is prudent to recognise and explain.

3.2 Summary

We found some literature discussing common barriers to using evidence to inform decisions. Four systematic reviews all identified the "communication and presentation" of evidence as a major barrier.

We found literature that addressed interventions to break down "communication and presentation" barriers and improve the use of evidence in decision making. It identified that user friendly design of evidence reports will improve the use of that evidence when making decisions.

We found no literature discussing the use of structured argumentation to improve decision making.

We found no literature that measured the impact of structured argumentation or evidence on decisions. We did find a study that looked to measure the impact of Health Technology Assessments

(HTA) on hospital decisions. They did this by comparing recommendations in the HTA to see if they were "consistent" with the hospital's decisions. This method is similar to our notion of "alignment" and thus goes some way to validating our methodology.

Although we found no literature that measured the impact of structured argumentation on policy decisions, there exists literature that argues for its positive impact on decision making and thinking in general. Most of the literature argues that a flexibly applied Toulmin argumentation scheme would help reports communicate their analysis better, and thus decision makers who read the reports would also benefit. We found literature establishing that training in argument mapping can substantially improve generic critical thinking skills. The structured argumentation in this literature resembles CASE in important respects.

3.3 Common barriers

The first question helps to frame the context of the study. Why might something like CASE be needed? If the literature suggests that common barriers to using evidence and arguments to inform decisions are barriers that CASE is designed to ameliorate, then CASE should help improve decision making.

Our review located several studies investigating the barriers to using evidence to inform decisions, though none dealt with biosecurity decision making.

(Lavis et al., 2005) conducted a systematic review of studies investigating decision making by health care managers/policy makers and conducted interviews with a sample of them to identify ways to improve the usefulness of systematic reviews for health care managers/policy makers. Of particular relevance are the interviews, which revealed a range of barriers to using evidence to inform decision-making. One such barrier was how the evidence is presented to the decision-maker. Specifically, interviewees stated that "most research reports are longer than can or will be read" and that "many health care managers and policy-makers actually read research reports [by] reading the abstract and conclusions first" (Lavis et al., 2005, p. 41). According to the authors, current approaches to presenting research evidence to decisions makers do not address these concerns. Perhaps more importantly, all 8 managers interviewed and 18/20 policy-makers, thought that "something like a 1:3:25 format (i.e. one page of take-home messages, a three-page executive summary, and a 25 page report)" would be an improvement on the current approach because the format is generally shorter than current approaches and "up-front placement of take-home messages reflects how many health care managers actually read reports" (Lavis et al., 2005, p. 41).

Another systematic review by (Mitton, Adair, Mckenzie, Patten, & Perry, 2007) focused on knowledge transfer between those who *produce* research and those who *use* that research to develop policy. Their review of the literature found a range of barriers to research evidence being used by decision makers. Of relevance to this study are the "communication" barriers that were identified, including: "Information overload and traditional academic language" (Mitton et al., 2007, p. 737). Similarly to the Lavis et al. review, they found that overcoming these barriers requires "clear summaries with policy recommendations" and a format that is "tailored to the specific audience" (Mitton et al., 2007, p. 737). Several of the studies reviewed claim that "each audience has different information needs and communication styles and therefore the information must be appropriately tailored" and "research should be presented in summary format, in simple language and with clearly worded recommendations" (Mitton et al., 2007, p. 738).

In a similar vein, (Orton, Lloyd-Williams, Taylor-Robinson, O'Flaherty, & Capewell, 2011) aimed to synthesise the evidence on the use of research by public health decision makers. The studies they

reviewed revealed a range of insights but of particular relevance is their identification of three studies which concluded that a major barrier to using evidence in decision making was “the structure of documents used to inform decisions” (Orton et al., 2011, p. 7).

(Poder, Bellemare, Bédard, Fisette, & Dagenais, 2018) looked at the impact Health Technology Assessments (HTA) have on local hospital decision making processes. Specifically, they sought to “identify the underlying factors for the non-implementation of recommendations”(Poder et al., 2018, p. 2). In other words, they wanted to identify what impediments exist to decision makers adopting the research findings of an HTA. Through structured interviews with hospital decision makers, they found that the third most frequently mentioned impediment to adopting HTA recommendations was content/format factors (14 percent) which includes the HTA report being poorly structured and not user-friendly as well as conclusions and recommendations of the report being imprecise (Poder et al., 2018, p. 4).

The three systematic reviews and interviews detailed above allow us to answer our first literature review question. A major barrier identified in them all is poor structure and presentation of evidence which impeded use of that evidence in decision making. Although these barriers were found in the health care domain, it is reasonable to suppose that they would also exist in a biosecurity context. Like health care, biosecurity decisions need to have a strong foundation in research evidence, some of which is complex and difficult to understand. Reports that present the research evidence to biosecurity decision makers would also benefit from what the systematic reviews identified, including in particular a standard format and up-front placement of conclusions. These are both facilitated by adopting CASE structure in reports.

3.4 Interventions

Now that communication barriers have been established as a major impediment to decision makers utilising evidence, the question then becomes *have any interventions been tested that seek to break down those barriers?*

(Langer, Tripney, & Gough, 2016) conducted two systematic reviews; the first review explored the evidence informed decision-making literature (EIDM) and the second looked at the social science literature on effective communication. The aim of the reviews was to explore “the efficacy of interventions applied to increase the use of research evidence by decision makers”(Langer et al., 2016, p. 13). Importantly, the systematic review is described by the authors as a review of reviews. That is, the systematic review only searched for and included other systematic reviews. For both reviews, they searched databases for keywords related to ‘research use’, hand searched key journals, checked reference lists of included studies and conducted forward citation checking exercises. For the second study the authors also used ‘snowball searches’, backward citation searches and introductory text-books.

Their reviews identified interventions that were grouped according to six mechanisms of change. These mechanisms of change are processes by which an increase in evidence use by decision makers may be achieved (Langer et al., 2016, p. 7). A summary of the interventions Langer et al. identified as having a positive effect on the use of evidence by decision makers and suggested means of implementing them are shown in the table below

Table 3-1: Summary of interventions and their means of implementation, from (Langer et al., 2016)

Intervention		Means of implementation
1	Awareness for, and positive attitudes towards, EIDM	1. Creation of behavioural norms around using evidence 2. Anchoring evidence use as a routine behaviour 3. Communicating risks of not using evidence 4. Social science principles for building user engagement
2	Agreement on policy relevant questions and fit-for-purpose evidence	1. Using consensus building techniques to determine what evidence is fit-for-purpose 2. Embed consensus building on fit-for-purpose evidence and policy relevance in wider efforts to build a professional identity
3	Communication of and access to evidence	1. Combine with motivation and opportunity building components 2. Evidence should be understandable and user-friendly 3. Tailoring and targeting evidence 4. Wording and contextualisation 5. Use of online and social media platforms
4	Facilitating interactions between decision makers and researchers	1. Use structured interactions 2. Fostering social influence engagement and sharing of norms and practices 3. Use online and mobile technologies 4. Improve understanding of decision makers' network structures
5	Skills to access and make sense of evidence	1. Combine with motivation and capability building components 2. Social science principles that enhance retention of learning results 3. Formally incorporate educational interventions into decision making structures and processes 4. Invest more efforts into building organisational and institutional EIDM capacities 5. Target thought processes and patterns rather than skill sets 6. Use online and mobile technologies to build capacity
6	Decision making processes and structures	1. Effective when incorporated with intervention 3 and 5 2. Use behavioural interventions to reduce cognitive biases and nudges to support the use of evidence 3. Direct facilitation of EIDM 4. Adopt organisation structures that are supportive of EIDM 5. Organisational mandate to support and institutionalise EIDM 6. Provide insights on decision-makers mental models, network structures, organisational settings and professional norms.

Of relevance to this study is the third mechanism identified "Communication and Access" which is described as the mechanism which "emphasises the importance of decision-makers receiving effective communication of evidence and convenient access to evidence"(Langer et al., 2016, p. 9). Out of the 36 systematic reviews reviewed, 11 dealt with this intervention specifically, and 32 dealt with it in combination with other interventions.

The specific components of the intervention mechanism are not all associated with the communication barriers identified in the preceding section. However, some of the components are relevant. In particular, the review of the EIDM literature found that "user friendly, hassle free design" of evidence products (Langer et al., 2016, p. 28) had a positive impact on the motivation of decision makers to use evidence. What made these evidence products more user friendly was having summary of findings table and plain language summaries. The broader social science literature reviewed found only one component of the mechanism that is relevant to our study. This was

"tailoring and targeting – aligning the communication of evidence to decision-makers professional needs and personal preference" (Langer et al., 2016, p. 29)

From the Langer et al. systematic review, we can conclude that those interventions that address communication barriers are found to be effective. We cannot make a direct comparison between the intervention discussed in the review and the CASE Structure because of differences in their nature. The intervention in the review was less focused on the format and presentation of evidence and did not explore the presentation of arguments at all. Furthermore, our study is interested in whether CASE adoption has improved decision making at MPI, whereas the intervention in the review is concerned with increasing the use of research evidence. Whilst it is generally agreed that increasing the use of evidence leads to better decisions, the review did not demonstrate this.

The point of including this systematic review in our own literature review is to show that interventions that improve how evidence is communicated to decisions makers has some impact. Since CASE adoption is also an intervention to improve how evidence and arguments are communicated, this lends plausibility to the conjecture that CASE adoption would also have some impact on decision making.

Apart from the systematic reviews referenced by the Langer et al. review, we were unable to locate any studies that tested the impact of interventions on the presentation and format of evidence/argument. This suggests a gap which our research will help address.

3.5 Methods for measuring the impact of reports on decision making

Another important topic in our review concerns methods devised to measure the impact that report recommendations have on decisions.

One study looked at the impact of Health Technology Assessments (HTA) on decision making in Austria. An HTA is a report designed to provide "independent and objective information on health technologies of various kinds for decision makers" (Zechmeister & Schumacher, 2012, p. 1). Their study sought to measure the impact HTAs had on decisions made by hospital financing boards. Part of the way they measured impact was to compare the recommendations in the HTA with the decision made by the hospital financing board for consistency. A decision is consistent when it adopts or implements the recommendations in the HTA. They found that in "45 percent (19) of the reports, the recommendation and decision were totally consistent. In 41 percent (17), technologies that had not been recommended were included on certain conditions, while in 12 percent (5), the decision was more restrictive than the recommendation" (Zechmeister & Schumacher, 2012, p. 5).

This relatively simple method of comparison hinges on the notion of "consistency". Decisions can be totally consistent, partially consistent or inconsistent with the recommendation in the HTA. When a decision is totally consistent with the HTA, the report has plausibly had some impact on that decision. Less so when partially consistent, and little impact at all when inconsistent. Consistency as a measure of impact is similar to our concept of alignment (Section 6).

There were limitations to using 'consistency' as a measure for the impact of HTA on decisions. Foremost, was accounting for influencing factors on decisions other than the HTA. Interviews with decision makers showed that organisational and administrative changes and regulations often had more impact on the decision makers than the HTA itself. Thus interviews with decision makers seem to be a crucial instrument for determining what influences decision making in this context, not just consistency between the HTA and the decision.

The authors also identify external sources of influence, such as changes in the perceived safety of the technology being assessed as well as negative public attitudes towards the technology. This external influence also contributes to whether or not the HTA is consistent with the hospital's decision. To fully understand the impact HTA have on decision making, these external influences would need to be factored in.

The need to understand and include factors other the HTA when analysing consistency between the document and the corresponding decision is relevant to Study 3: Are post-CASE IHS decisions better aligned with risk assessments? While the IRA is used to inform risk management decisions, it is not the sole source of influence. To fully understand the extent to which sources other than the IRA influence decisions, we too would need to conduct interviews or surveys with MPI decision makers. This is discussed further in Section 7.3.3, New research directions.

We reviewed other studies by (Dannenberg, 2016; Gagnon, Desmartis, Poder, & Witteman, 2014; Poder et al., 2018) which also looked to measure the impact of health reports on decision makers. The methods used to measure those impacts were either surveys, interviews, literature reviews or simply the authors' judgements. While this project does not use these methods, the surveys and interviews conducted with decision makers in these studies could be used to inform future research. Of particular usefulness is the survey instrument developed by (Poder et al., 2018, p. 3) used a scale to measure the impact that recommendations in reports have on hospital decision makers:

Figure 3-1: Levels of impact of recommendations from (Poder et al., 2018).

1. **Awareness**: the report is known and has been consulted
2. **Acceptance**: the report is relevant and perceived as useful for decision making
3. **Appropriation**: recommendations are explicitly considered by decision makers
4. **Adoption**: recommendations are adopted
5. **Implementation**: recommendations are implemented
6. **Outcomes**:
    - 6.1. Used as reference material: the report and recommendations are used by clinicians and managers and are considered as a reference to help better practice
    - 6.2. Inclusion in policies or administrative documents: the report and recommendations are incorporated in the documentation used by clinicians and managers
    - 6.3. Practitioners' conditions of practice: the report and recommendations have an impact on working conditions
    - 6.4. Change in practice: the report and recommendations lead to changes in clinical and/or managerial practices
    - 6.5. Efficient use of the resources: a better value for money is provided (i.e., cost reduction or improved outcome)
    - 6.6. Impact on patients: the health status and/or wellbeing of patients is improved

The 'Outcomes' impacts are too specific to healthcare to be used directly in the biosecurity context, however with some modification they could find relevance. Moreover, Levels 1-5 seem generally applicable and would work well to capture how IRA reports influence decision making at MPI. We include this brief discussion of the (Poder et al., 2018) survey instrument to demonstrate how decision makers are included in research on their decisions. We discuss doing the same in future research in Section 7.3.

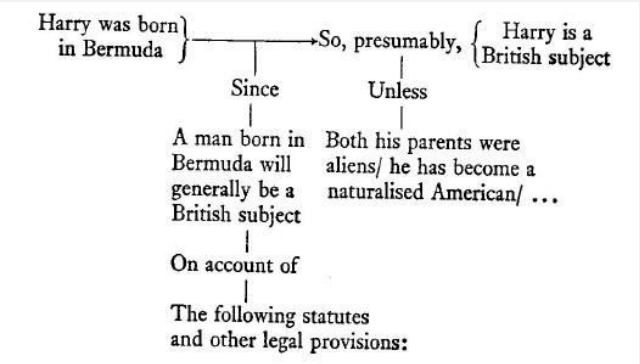
3.6 Models of structured argumentation and impact on decisions

Because an objective of this study was to measure the impact that CASE adoption at the MPI had on their biosecurity decisions, and CASE is a kind of structured argumentation, any literature on the impact that structured argumentation has on policy decision making would be relevant. However, we were unable to locate any studies directly measuring such impact. We were able to locate literature arguing that structured argumentation *should* be beneficial, as well as literature on the impact of training in structured argumentation on critical thinking skills.

Structured argumentation has traditionally been based on philosopher Stephen Toulmin's model of argumentation. The Toulmin model has 6 components (Toulmin, 2003, pp. 87–100):

1. Claim (C) – What is being argued for, i.e. what one is attempting to establish as true or false.
2. Data (D) – The facts we appeal to establish the claim.
3. Warrant (W) – The inferential leap that takes one from D to C.
4. Backing (B) – Whatever adds credibility or authority to the warrant.
5. Quantifier (Q) – Indicates the strength conferred to C by W and B.
6. Rebuttal (R) – Circumstances where the authority of the warrant would have to be set aside.

Figure 3-2: The Toulmin model of argumentation. Diagram from (Toulmin, 2003).



It is worthwhile noting some similarities and differences between the Toulmin model of argument and CASE (see Section 2.3.1 below for description of CASE). A main contention is one of the primary features of a CASE Structured argument. It is defined as a "contestable proposition" (van Gelder, 2019, p. 11), where "contestable" means that there is room for reasonable people to disagree, and a proposition is a statement that is either true or false The idea is a familiar one within argument theory and informal logic. In Toulmin's argument model, instead of calling it a *contention* he calls it a *claim* which is "the conclusion of the argument and the point at issue in a controversy" (Kneupper, 1978, p. 238). Similar to the CASE Structure, the conclusion is established by data and warrants whereas in a CASE structured argument, the contention is established by arguments and evidence.

The components of a Toulmin argument have been applied in a variety of domains. Examples include: to build confidence in decisions made by social workers (Duffy, 2011), to structure the arguments in digital forensic practice (Franqueira & Horsman, 2020), to build safety cases (Bloomfield & Bishop, 2010) and for security requirements analysis (Haley, Laney, Moffett, & Nuseibeh, 2008). The application in these domains is argued for through worked examples and case studies with the aim of demonstrating why it is, or ought to be, beneficial to the domain. These benefits are summarised below:

- When making decisions that need to balance competing interests, a focus on justifying the warrant in a systematic and structured way builds confidence in the decision (Duffy, 2011).
- Using structured argumentation presents the argument in a readable and accessible way (Franqueira & Horsman, 2020; Kelly, 2004)

- Using structured argumentation allows readers to reconstruct or trace how a conclusion was reached (Franqueira & Horsman, 2020)
- Constructing an structured argument, particularly with rebuttals, helps identify vulnerabilities in a system (Haley et al., 2008)

This literature shows that structured argumentation is being adopted in a variety of domains, with presumed benefits. However, none of the reviewed literature showed that its adoption has actually brought about such benefits.

Structured argumentation is said to have beneficial effects to thinking and writing in general. In business and management consulting, the most widely applied structure in this regard is the Pyramid Principle (Minto, 2009). Minto argues that one should think through and present ideas to a reader from the 'top down' in a pyramidal structure. To order ideas this way one must obey three rules (Minto, 2009, p. 9):

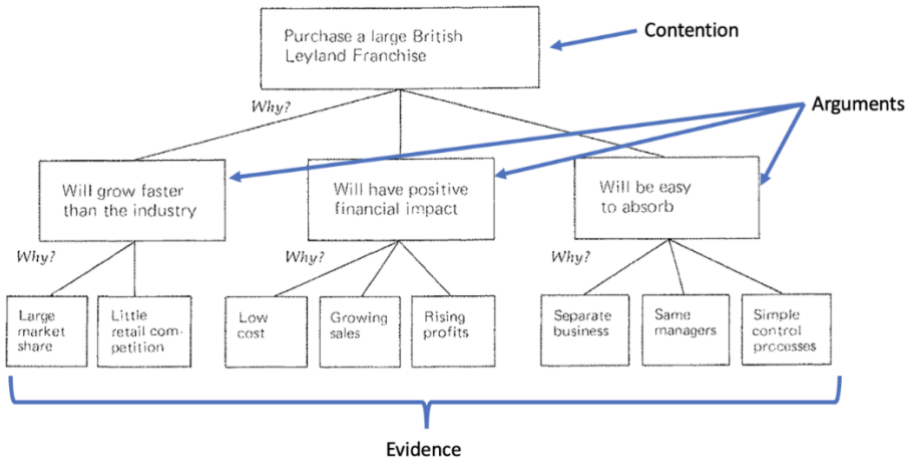
1. Ideas at any level in the pyramid must always be summaries of the ideas grouped below them – the major activity carried out in thinking and writing is that of abstracting to create a new idea out of the ideas grouped below.
2. Ideas in each grouping must always be the same kind of idea – each item in the group needs to be at the same level of abstraction.
3. Ideas in each grouping must always be logically ordered – either ordered deductively, chronologically, structurally or comparatively.

Following these rules results in a pyramid of ideas that can then be translated to prose in a document. Minto gives the example of a purchasing a large British Leyland Franchise (Figure 3-3).

According to Minto, the benefit of organising reasoning in this way is that it "forces visual recognition of the question/answer relationship on you as you work out your thinking. Any point you make must raise a question in the reader's mind, which you must answer on the line below" (p.16). Once the pyramid is constructed, it can be translated to the page with formatting matching the level of abstraction in the pyramid. In this example, the title or heading of the page would be "Purchase a Large British Leyland Franchise" followed by three section headings "will grow faster than the industry", "will have positive financial impact" and "will be easy to absorb". Next, subsection headings would represent each of the lower-level ideas.

There is a striking similarity between the Pyramid Principle and CASE structure. With the Pyramid Principle, each level is an abstraction of the one below and likewise with CASE, the idea is that "in a given line of reasoning, the Evidence and Arguments incrementally increase in generality, from the tiniest details up to the most general points" (van Gelder, n.d., p. 33). In fact, we could take the example pyramid above and label it with the elements of CASE (excluding sourcing):

Figure 3-3: Example of a Minto-style pyramid of ideas annotated with CASE terminology.



The labelling illustrates the CASE principle of leveraging the "ladder of abstraction" (van Gelder, n.d., p. 26).

The Pyramid Principle was included in this review to illustrate that structuring arguments in a way similar to CASE is widely believed to help organise and clearly present the author's thinking. Given the wide use of the Pyramid Principle in management consulting, it stands to reason that CASE, being similar in relevant respects, might be equally beneficial. However we were unable to find any scientific research evaluating the extent to which the Pyramid Principle leads to better quality reports or decision making.

Finally we consider studies that sought to measure the impact of argument mapping training on critical thinking skills. Intensive training in argument mapping as a means to develop critical thinking skills, and the systematic evaluation of that training using pre- and post-testing with an objective[10] critical skills test and a control condition, was first introduced at the University of Melbourne in the late 1990s (Rider & Thomason, 2008; Twardy, 2004). A series of studies found that this approach produced substantial gains in critical thinking skills (van Gelder et al., 2004). A meta-analysis of the Melbourne studies, plus studies from other institutions, found a large pre-post effect size for this technique (Alvarez, 2007). This effect has subsequently been replicated many times using a variety of critical thinking skills tests – e.g. (Cullen, Fan, van der Brugge, & Elga, 2018; Thomason, 2014).

We can thus be confident that intensive training in argument mapping improves general critical thinking skills. The CASE training used at MPI is a version of argument mapping training, quite closely based on the training used in the studies described above, albeit not as intensive. It is thus plausible that the MPI CASE training had at least some benefit for participants' general critical thinking skills, which may have contributed to improved reasoning in MPI reports alongside the impact of directly applying the specific CASE techniques,[11] which may in turn have contributed to better decisions.

[10] "Objective" here means a test which can be scored without expert human judgement (i.e., in practice, a multi-choice test) and which was developed independently of the research in which it is applied.

[11] In this regard the Cullen et. al. study is particularly relevant. One of their findings was that Princeton students who had had intensive argument mapping training turned in far better-reasoned essays than comparable students who had not had the training.

4 Study 1: Do post-CASE IRAs show stronger CASE structure?

As described in the Introduction, we resolved the overarching question for his project – is the CASE initiative helping improve decision making at MPI? – into two main research questions:

1. To what extent has the CASE initiative improved the clarity and rigour of reasoning in IRAs?
2. To what extent has this improvement (if any) led to better decisions?

Our first study addresses the first question. We focus on whether the CASE initiative has succeeded in its most immediate aim: that reasoning in IRAs should be improved in the sense of better reflecting CASE principles.

4.1 Objective

Estimate the extent to which post-CASE IRAs exhibit stronger CASE structure than pre-CASE IRAs.

4.2 Method

4.2.1 Overview

Study 1 used a retrospective observational design. We took a sample of IRAs from before, and after, the start of the CASE initiative in 2015, and sampled sections of reasoning from within these IRAs. We developed a framework for coding these sections for the presence of CASE structure. We applied the coding framework to the reasoning samples, and used statistical analysis to understand the difference between pre- and post-CASE samples.

4.2.2 Sampling

In principle, it would be possible to compare all reasoning in all pre- and post-CASE IRAs. However this was not feasible given the resources available for this project, and various other constraints. Instead, we took samples, taking care to ensure that the pre- and post- samples were matched in relevant respects. This sampling took place at two levels. First, we took samples of IRA reports; then we took samples of sections of reasoning from within those reports.

4.2.2.1 Sampling IRAs

With the help of an expert in MPI, we constructed a purposeful sample of Plants division IRAs. Purposeful sampling, unlike sampling based on purely statistical considerations, involves selecting the sample to achieve particular objectives and satisfy relevant constraints. It can be appropriate for qualitative research, or (as in this case) exploratory quantitative research (Patton, 2015). Note that purposeful sampling can include representativeness as one objective among others.

Our objectives in our purposeful sampling of IRAs were:

- Make the sample as representative as we could, given other constraints;
- Ensure that the sub-sample of pre-CASE IRAs was similar in key respects to the sub-sample of post-CASE IRAs, such as being matched in the commodity types they were assessing;
- Ensure that the selection of post-CASE IRAs provides a suitable opportunity to identify the phenomenon of interest, viz., stronger CASE structure, if it is present. (This potentially conflicts with representativeness.)
- Keep the sample to a manageable size, given our resources.

Table 4-1: The 18 IRAs received from MPI and their capacity to be coded. IRAs not excluded by one of the constraints are designated by shading.

	Year	IRA	Commodity Type	Reason for exclusion
Pre-CASE				
1	2007	Miscanthus plants in vitro from UK and USA	Nursery Stock	Potentially sensitive
2	2007	Litchi chinensis fresh fruit from Taiwan	Fresh Fruit/Vegetables	No IHS available
3	2007	Vehicles and Machinery	Vehicle and Machinery	No corresponding post-CASE IRA of same commodity type.
4	2008	Wollemi Pine Nursery Stock	Nursery Stock	Potentially sensitive
5	2008	Litchi fresh fruit from Australia	Fresh Fruit/Vegetables	
6	2009	Table grapes from China	Fresh Fruit/Vegetables	
7	2009	Onion Fresh Bulbs from China	Fresh Fruit/Vegetables	Potentially sensitive
8	2009	Fresh Coconut from Tuvalu	Fresh Fruit/Vegetables	
9	2009	Pears from China	Fresh Fruit/Vegetables	
10	2008	Fresh Island Cabbage Leaves	Fresh Fruit/Vegetables	Not enough information in IHS
11	2012	Inorganic Fertiliser	Fertiliser	No corresponding post-CASE IRA of same commodity type; Not enough information in IHS
12	2012	Tomato and Capsicum Seeds for Sowing	Seeds for Sowing	No corresponding post-CASE IRA of same commodity type; Not enough information in IHS
13	2012	Malus nursery stock from all countries	Nursery Stock	
14	2013	Rosa Nursery Stock	Nursery Stock	IRA does not separate analysis of risk for different stages of growth whereas IHS does. This makes alignment difficult to reconcile.
Post-CASE				
15	2016	Fresh salacca fruit from Indonesia	Fresh Fruit/Vegetables	IHS still in draft
16	2016	Fresh Rambutan from Vietnam	Fresh Fruit/Vegetables	
17	2018	Actinidia Plants for Planting	Nursery Stock	
18	2019	Prunus Plants for Planting	Nursery Stock	

We identified a total of 18 IRAs produced in Plants and Fresh Fruit/Vegetables divisions over the period 2007-2019, of which 14 were pre-CASE and 4 were post-CASE (Table 4-1). Of these, many were unsuitable, and so excluded, for a variety of reasons. These reasons included:

- Some IRAs were potentially sensitive in that the author(s) had not given explicit consent to having them analysed for clarity and rigour of reasoning.
- Some pre-CASE IRAs could not by matched by a post-CASE IRA of the same commodity type, or vice versa.
- Some IRAs did not have a corresponding IHS, or the IHS was not suitable for coding. We excluded these in preparation for Study 3, in which we would be working with IRA/IHS pairs.

It should also be noted here that out of all reports we received from MPI, only one - Prunus for Planting - explicitly stated that analyses would be structured using CASE (Berry, Durrant, Narouei Khandan, Wilson, & Philip, 2019, p. 28)

After exclusions we were left with three post-CASE IRAs of two types – Fresh Fruit/Vegetables, and Nursery Stock. To have our pre- and post-CASE samples be as similar as possible, each of these should be paired with a pre-CASE IRA of the same type and as similar as possible in other respects such as length. We paired Fresh Rambutan from Vietnam (post) with Pears from China (pre). There were two post-CASE Nursery Stock IRAs, but only one non-excluded pre-CASE Nursery Stock IRA (Malus). Of the two post-CASE IRAs, we chose Prunus rather than Actinidia for pairing with Malus because Prunus had been explicitly structured with CASE, and so gave the best opportunity for the impact of CASE of the CASE initiative to be manifested. The resulting samples are summarised in Table 4-2.

Table 4-2: Final sample of IRAs for Study 1. We used all non-excluded post-CASE IRAs which could be paired with a pre-CASE IRA of the same commodity type.

Commodity Type	**Pre**	**Post**
Fresh Fruit/Vegetables	Pears from China (2009)[12]	Rambutan from Vietnam (2016)[13]
Nursery Stock	Malus nursery stock (2012)[14]	Prunus for Planting (2019)[15]

4.2.2.2 Sampling sections of reasoning

From the selected IRA in our higher-level sample, we selected specific sections of reasoning upon which we could apply our coding scheme. Each IRA assessed is composed of similar subsections. For

[12] Tyson, Joy, Selma Rainey, Joan Breach, and Sandy Toy. "Import Risk Analysis: Pears (Pyrus Bretschneideri, Pyrus Pyrifolia, and Pyrus Sp. Nr. Communis) Fresh Fruit from China." Ministry for Primary Industries, New Zealand, 2009. https://www.mpi.govt.nz/dmsdocument/2884-Pears-Pyrus-bretschneideri-Pyrus-pyrifolia-and-Pyrus-sp.-nr.-communis-fresh-fruit-from-China-Final-Risk-Analysis-October-2009.

[13] Clark, Sarah. "Import Risk Analysis: Fresh Rambutan from Vietnam." Ministry for Primary Industries, New Zealand, 2016. https://www.mpi.govt.nz/dmsdocument/14254-Import-Risk-Analysis-Fresh-Rambutan-from-Vietnam.

[14] Ormsby, Mike, and Lihong Zhu. "Import Risk Analysis: Viruses, Viroids, Phytoplasma, Bacteria and Diseases of Unknown Aetiology on Malus Nursery Stock from All Countries." Ministry for Primary Industries, New Zealand, 2012. https://www.mpi.govt.nz/dmsdocument/2873-Apple-Malus-domestica-nursery-stock-micro-organisms-and-diseases-Final-Import-Risk-Analysis-July-2012.

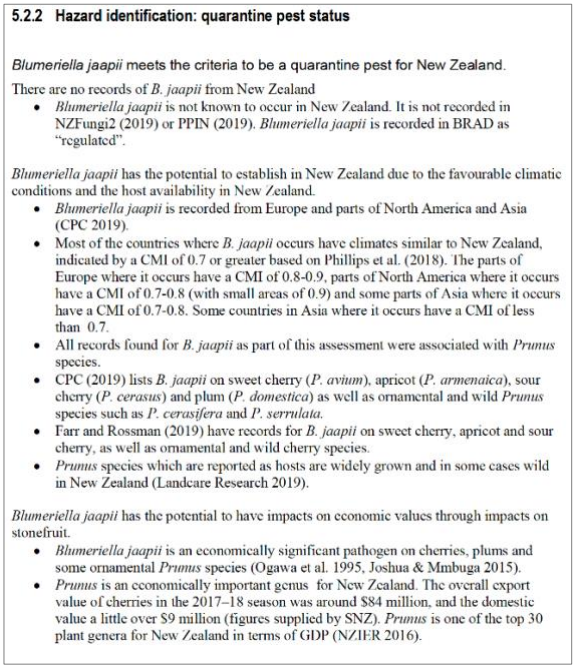
[15] Berry, Jocelyn A, Abigail Durrant, Hossein Narouei Khandan, Karen Wilson, and Bruce Philip. "Import Risk Analysis: Prunus Plants for Planting." Ministry for Primary Industries, New Zealand, 2019. https://www.mpi.govt.nz/dmsdocument/38693-20191125-Version-2-Final-Draft-Import-Health-Standard-for-Prunus-Plants-for-Planting.

example, in the Pears from China IRA, there are 75 “Economic Consequences” subsections, 72 “Entry Assessments” subsections, 69 “Establishment Assessment” subsections and so on.

Most of the subsections can be analysed as arguments because they attempt to advance a claim or draw a conclusion and cite evidence in support of it. Thus, these subsections form a natural unit of analysis onto which the coding scheme could be applied. We termed these subsections “Component Arguments” (CA).

Coding the entirety of each IRA was not necessary. CA types are repeated throughout the report and within a report only differ in their subject matter, not their structure. Take for example the following two “Hazard Identification: Quarantine Pest Status” CAs from the Prunus Plants for Planting IRA:

Figure 4-1: Hazard identification: quarantine pest status for *Blumeriella jaapii* (Berry et al., 2019, p. 37) (left) and *Phytophthora palmivora* (right)



5.2.2 Hazard identification: quarantine pest status

Blumeriella jaapii meets the criteria to be a quarantine pest for New Zealand.

There are no records of *B. jaapii* from New Zealand

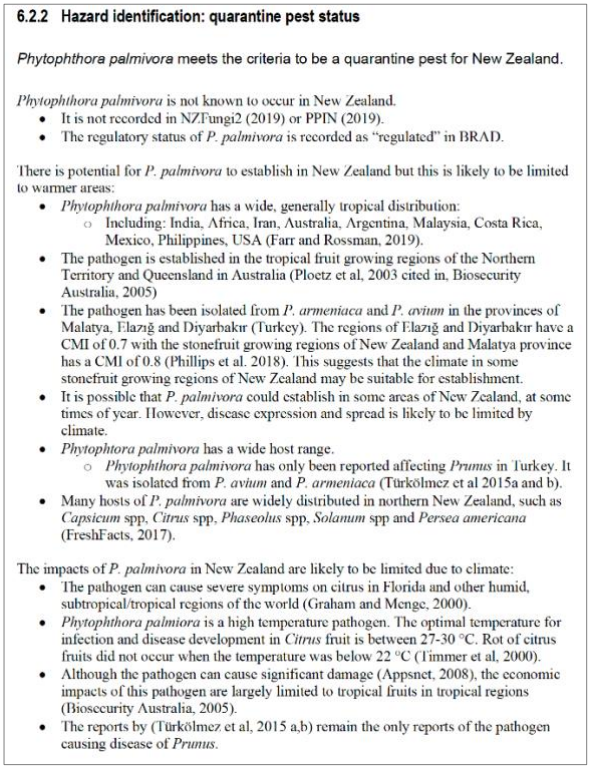
- *Blumeriella jaapii* is not known to occur in New Zealand. It is not recorded in NZFungi2 (2019) or PPIN (2019). *Blumeriella jaapii* is recorded in BRAD as “regulated”.

Blumeriella jaapii has the potential to establish in New Zealand due to the favourable climatic conditions and the host availability in New Zealand.

- *Blumeriella jaapii* is recorded from Europe and parts of North America and Asia (CPC 2019).
- Most of the countries where *B. jaapii* occurs have climates similar to New Zealand, indicated by a CMI of 0.7 or greater based on Phillips et al. (2018). The parts of Europe where it occurs have a CMI of 0.8-0.9, parts of North America where it occurs have a CMI of 0.7-0.8 (with small areas of 0.9) and some parts of Asia where it occurs have a CMI of 0.7-0.8. Some countries in Asia where it occurs have a CMI of less than 0.7.
- All records found for *B. jaapii* as part of this assessment were associated with *Prunus* species.
- CPC (2019) lists *B. jaapii* on sweet cherry (*P. avium*), apricot (*P. armenaica*), sour cherry (*P. cerasus*) and plum (*P. domestica*) as well as ornamental and wild *Prunus* species such as *P. cerasifera* and *P. serrulata.*
- Farr and Rossman (2019) have records for *B. jaapii* on sweet cherry, apricot and sour cherry, as well as ornamental and wild cherry species.
- *Prunus* species which are reported as hosts are widely grown and in some cases wild in New Zealand (Landcare Research 2019).

Blumeriella jaapii has the potential to have impacts on economic values through impacts on stonefruit.

- *Blumeriella jaapii* is an economically significant pathogen on cherries, plums and some ornamental *Prunus* species (Ogawa et al. 1995, Joshua & Mmbuga 2015).
- *Prunus* is an economically important genus for New Zealand. The overall export value of cherries in the 2017–18 season was around $84 million, and the domestic value a little over $9 million (figures supplied by SNZ). *Prunus* is one of the top 30 plant genera for New Zealand in terms of GDP (NZIER 2016).



6.2.2 Hazard identification: quarantine pest status

Phytophthora palmivora meets the criteria to be a quarantine pest for New Zealand.

Phytophthora palmivora is not known to occur in New Zealand.

- It is not recorded in NZFungi2 (2019) or PPIN (2019).
- The regulatory status of *P. palmivora* is recorded as “regulated” in BRAD.

There is potential for *P. palmivora* to establish in New Zealand but this is likely to be limited to warmer areas:

- *Phytophthora palmivora* has a wide, generally tropical distribution:
  - Including: India, Africa, Iran, Australia, Argentina, Malaysia, Costa Rica, Mexico, Philippines, USA (Farr and Rossman, 2019).
- The pathogen is established in the tropical fruit growing regions of the Northern Territory and Queensland in Australia (Ploetz et al, 2003 cited in, Biosecurity Australia, 2005)
- The pathogen has been isolated from *P. armeniaca* and *P. avium* in the provinces of Malatya, Elazığ and Diyarbakır (Turkey). The regions of Elazığ and Diyarbakır have a CMI of 0.7 with the stonefruit growing regions of New Zealand and Malatya province has a CMI of 0.8 (Phillips et al. 2018). This suggests that the climate in some stonefruit growing regions of New Zealand may be suitable for establishment.
- It is possible that *P. palmivora* could establish in some areas of New Zealand, at some times of year. However, disease expression and spread is likely to be limited by climate.
- *Phytophtora palmivora* has a wide host range.
  - *Phytophthora palmivora* has only been reported affecting *Prunus* in Turkey. It was isolated from *P. avium* and *P. armeniaca* (Türkölmez et al 2015a and b).
- Many hosts of *P. palmivora* are widely distributed in northern New Zealand, such as *Capsicum* spp, *Citrus* spp, *Phaseolus* spp, *Solanum* spp and *Persea americana* (FreshFacts, 2017).

The impacts of *P. palmivora* in New Zealand are likely to be limited due to climate:

- The pathogen can cause severe symptoms on citrus in Florida and other humid, subtropical/tropical regions of the world (Graham and Menge, 2000).
- *Phytophthora palmiora* is a high temperature pathogen. The optimal temperature for infection and disease development in *Citrus* fruit is between 27-30 °C. Rot of citrus fruits did not occur when the temperature was below 22 °C (Timmer et al, 2000).
- Although the pathogen can cause significant damage (Appsnet, 2008), the economic impacts of this pathogen are largely limited to tropical fruits in tropical regions (Biosecurity Australia, 2005).
- The reports by (Türkölmez et al, 2015 a,b) remain the only reports of the pathogen causing disease of *Prunus*.

We can see that while CAs differ in the pests they deal with, they have very similar structure. Both consider three criteria that a pest must meet to be classified as a hazard. While the number of criteria that the pest is evaluated against may change, the argumentative structure does not. That is, for this type of CA, pests are always assessed against criteria to determine if they are a hazard. Thus, a representative sample of the CA types would be sufficient to determine the extent to which the IRA instantiates the CASE structure.

To assemble this representative sample of CAs, we determined which CA types appeared most frequently in the report. We then randomly chose up to 6 CA from those types until we reached a maximum of 25 (See CA types coded and their frequency for more details)

In the case of Rambutan from Vietnam, there were only 17 CAs in the whole document so we coded them all. We coded a maximum of 25 CAs from each IRA because that was the upper limit of what was manageable given our resources.

The final sample of CAs is shown in the table below.

Table 4-3: Final sample of CAs drawn from the selected IRAs.

IRA	Pre or Post CASE	Number of CAs
Pears from China	Pre	25
Malus Nursery Stock	Pre	25
Rambutan from Vietnam	Post	17
Prunus Plants for Planting	Post	25
	Total	92

4.2.3 Procedure

4.2.3.1 Developing a coding scheme

We developed a coding scheme in a question-and-answer format. The scheme asked nine questions that would be answered one way if CASE structure was present and another if it was not. In the initial version, the questions were binary (yes/no). However, in a pilot of the scheme, we found that each question admitted a range of possible answers, depending on the degree to which CASE was present in the CA. We therefore revised the coding scheme to ensure it captured the "qualitative richness of the phenomenon" (Boyatzis, 1998, p. 1), where "the phenomenon" in our study is the structure and presentation of the CAs. A full list of the resulting questions, answers and their corresponding scores can be found in Appendix 1.

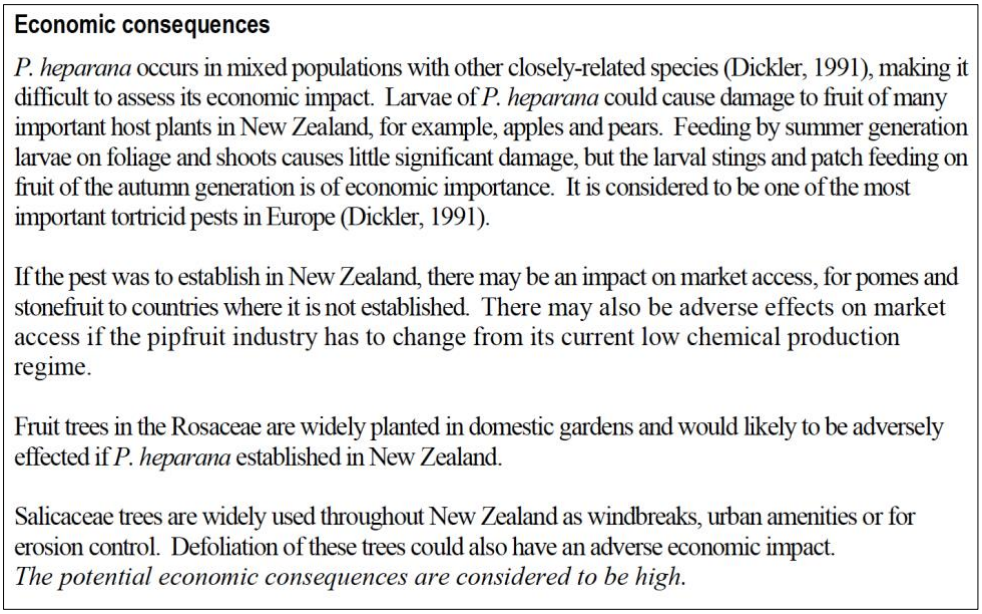
4.2.3.2 Applying the coding scheme

After our pilot coding we applied the revised scheme to our full sample of CA..

4.2.3.2.1 Example of coding a CA

To illustrate the process, consider the following CA where the economic consequences of a particular pathogen, *P. heparana*, establishing in NZ are assessed.

Figure 4-2: Screenshot of a CA from an MPI IRA. The CA assesses the potential economic consequences of *P. heparana* (Tyson, Rainey, Breach, & Toy, 2009, p. 265).



Economic consequences

P. heparana occurs in mixed populations with other closely-related species (Dickler, 1991), making it difficult to assess its economic impact. Larvae of *P. heparana* could cause damage to fruit of many important host plants in New Zealand, for example, apples and pears. Feeding by summer generation larvae on foliage and shoots causes little significant damage, but the larval stings and patch feeding on fruit of the autumn generation is of economic importance. It is considered to be one of the most important tortricid pests in Europe (Dickler, 1991).

If the pest was to establish in New Zealand, there may be an impact on market access, for pomes and stonefruit to countries where it is not established. There may also be adverse effects on market access if the pipfruit industry has to change from its current low chemical production regime.

Fruit trees in the Rosaceae are widely planted in domestic gardens and would likely to be adversely effected if *P. heparana* established in New Zealand.

Salicaceae trees are widely used throughout New Zealand as windbreaks, urban amenities or for erosion control. Defoliation of these trees could also have an adverse economic impact.
The potential economic consequences are considered to be high.

This CA would be coded as follows:

Table 4-4: Coding example with justifications.

	Question	Answer	**Score**	Justification
1.	Is there a main contention?	Yes	1	The main contention is italicized and presented at the end of the CA
2.	Is the main contention easily identifiable as such?	Yes	1	The main contention is easy to identify as it is separated from the body of the text and italicized.
3.	Is the main contention positioned at the top?	No	0	The main contention is at the end of the CA
4.	Is there a high-level argument structure?	Partially	0.25	There is a trace of a high-level argument but mostly the CA reads like collection of evidence.
5.	Is the high-level argument structure easily identifiable as such?	No	0	The trace of high-level argument is very difficult to discern. No attempt has been made to delineate what the reasons are for believing the contention.
6.	If there are high-level arguments, is it clear which items of evidence, if any, are intended to support the arguments?	Partially	0.25	Since the reasons are difficult to discern, it's also difficult to tell which items of evidence are supposed to support them. In the first paragraph it is clear that "larval stings and patch feeding on fruit" is a piece of evidence intended to support "Larvae of P. heparana could cause damage to fruit". Thus, for this question, "partially" is appropriate.
7.	Are items of evidence clearly relevant to the reasons?	Partially	0.25	While some items of evidence are relevant, for others it is hard to tell. For instance, how is a "change from its current low chemical production regime" related to the reason "there may also be adverse effects on market access". If this evidence is relevant it's certainly not clear. Of course, experts in domain might be able to fill in the link with their domain knowledge. However, to the general reader the link between the two items of evidence is unclear.
8.	Are items of evidence presented as such?	No	0	No attempt has been made to demarcate what is evidence and what isn't.
9.	Is each item of evidence appropriately sourced?	Partially	0.25	Some items of evidence are sourced, but others are not. Therefore "partially" is warranted.
Total			3/9	

In this way we are able to generate a score for each CA out of 9. We then averaged the scores for all CAs from an IRA to get an overall picture of the degree to which CASE has been adopted in that IRA.

4.2.3.2.2 Steps taken to avoid bias

With any qualitative content analysis, there is potential for bias to influence the results. For this study, there is potential for 1) observer bias and 2) selection bias. Observer bias is any kind of deviation from the truth during the process of observing and recording information due to the

perception of the observer (Mahtani, Spencer, Brassey, & Heneghan, 2018) in our case, the coder. Typically, to reduce the effects of this bias, multiple coders are used, the idea being that if multiple coders agree despite the biases peculiar to each, then it's likely the are picking up on a phenomenon in the data rather than an artefact of their perception. However, this study was a pilot attempting to explore a methodology that would meet the studies objectives. As a pilot, resources were more constrained, and results were not expected to be as conclusive as they might be with a more rigorously structured study design. What's more, coding the CA takes a certain degree of expertise in structured argumentation. Finding multiple coders with that expertise was not feasible. We therefore only used one coder. We did make an attempt to blind the coder to the source of the CA so they wouldn't know if it was an example of a pre-CASE CA or a post-CASE CA, another common strategy to mitigate observer bias. However, we quickly realised that this effort was futile as there was too much information specific to the source of the CA contained therein. To summarise, we were unable to mitigate the effect of observer bias for this study.

Selection bias occurs when the sample being studied is biased in some way that makes it unrepresentative of the population. To avoid this bias, the CAs were randomly selected from the most frequently instantiated types of CA. To ensure randomisation, we assigned each CA from selected types a random number using the rand() function in excel, sorted the column from smallest to largest and then incorporated the first 3-5 in our sample. Potentially, selecting from the most frequently instantiated types might introduce a selection effect, but if we included less frequently used types the sample would not reflect the bulk of the document. We opted for greater coverage of the documents we were coding by selecting CAs from the most frequently used types.

4.3 Results

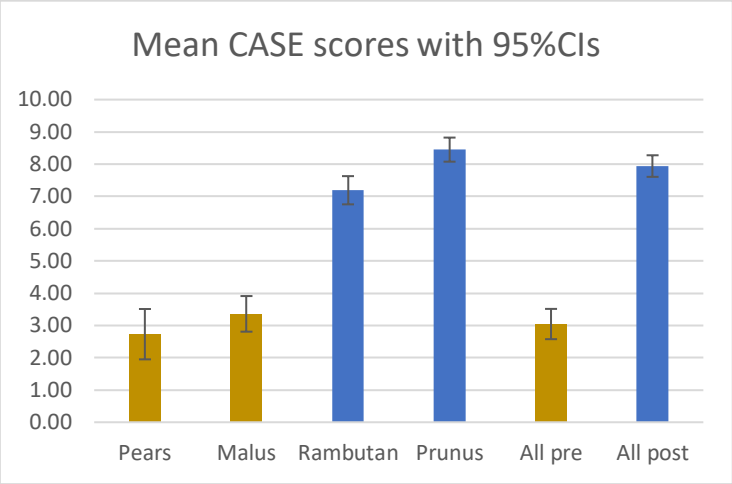
4.3.1 Summary statistics

The coding scheme we applied generated an average pre-CASE score of 3.05 and a post-CASE score of 7.94.

Table 4-5: Descriptive statistics of coding results.

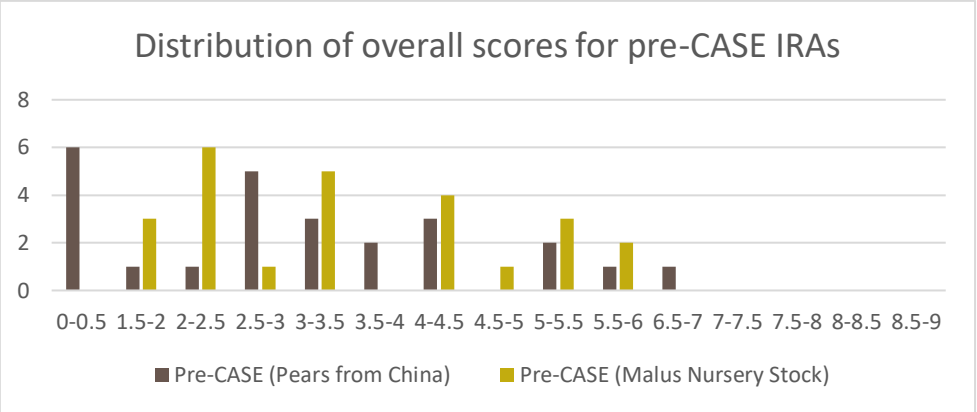
IRA	Pre or Post CASE	n	Mean	95% CI on mean	Median	SD	Standard Error
2009 – Pears from China	Pre	25	2.73	(1.95, 3.51)	2.75	1.89	0.38
2012 – Malus Nursery Stock	Pre	25	3.36	(2.81, 3.91)	3	1.34	0.27
2016 – Rambutan	Post	17	7.19	(6.75, 7.63)	7.5	0.85	0.21
2019 – Prunus for Planting	Post	25	8.45	(8.08, 8.82)	8.75	0.9	0.18
	All pre-CASE	50	3.05	(2.58, 3.51)	3	1.65	0.23
	All post-CASE	42	7.94	(7.61, 8.27)	8	1.07	0.17

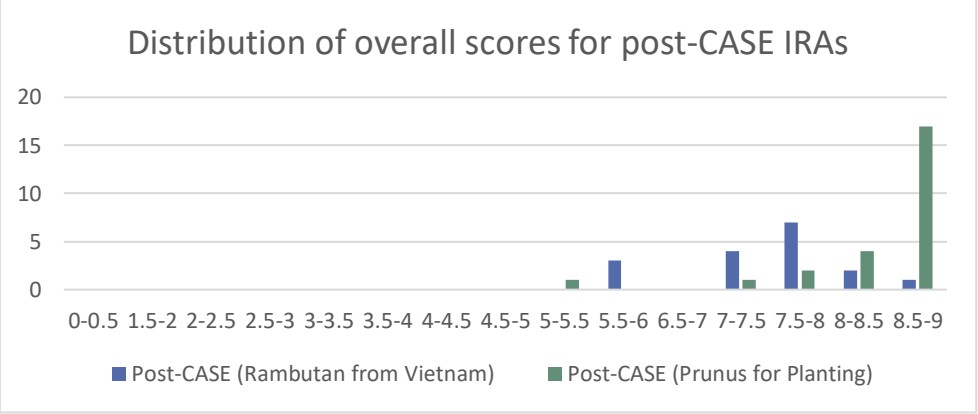
Figure 4-3: Mean CASE scores for CAs drawn from reports before (Pears, Malus), and after (Rambutan, Prunus), the start of CASE training, with 95% confidence intervals.[16]



For a little more insight into the improvement from pre- to post-CASE training, we can compare the distribution of overall scores in pre-CASE reports versus post-CASE reports.

Figure 4-4: Distribution of overall scores for component arguments from pre-CASE IRAs (top) and post-CASE IRAs (bottom). Each bar represents the number of component arguments from a given IRA whose overall score is in the range shown on the x-axis.





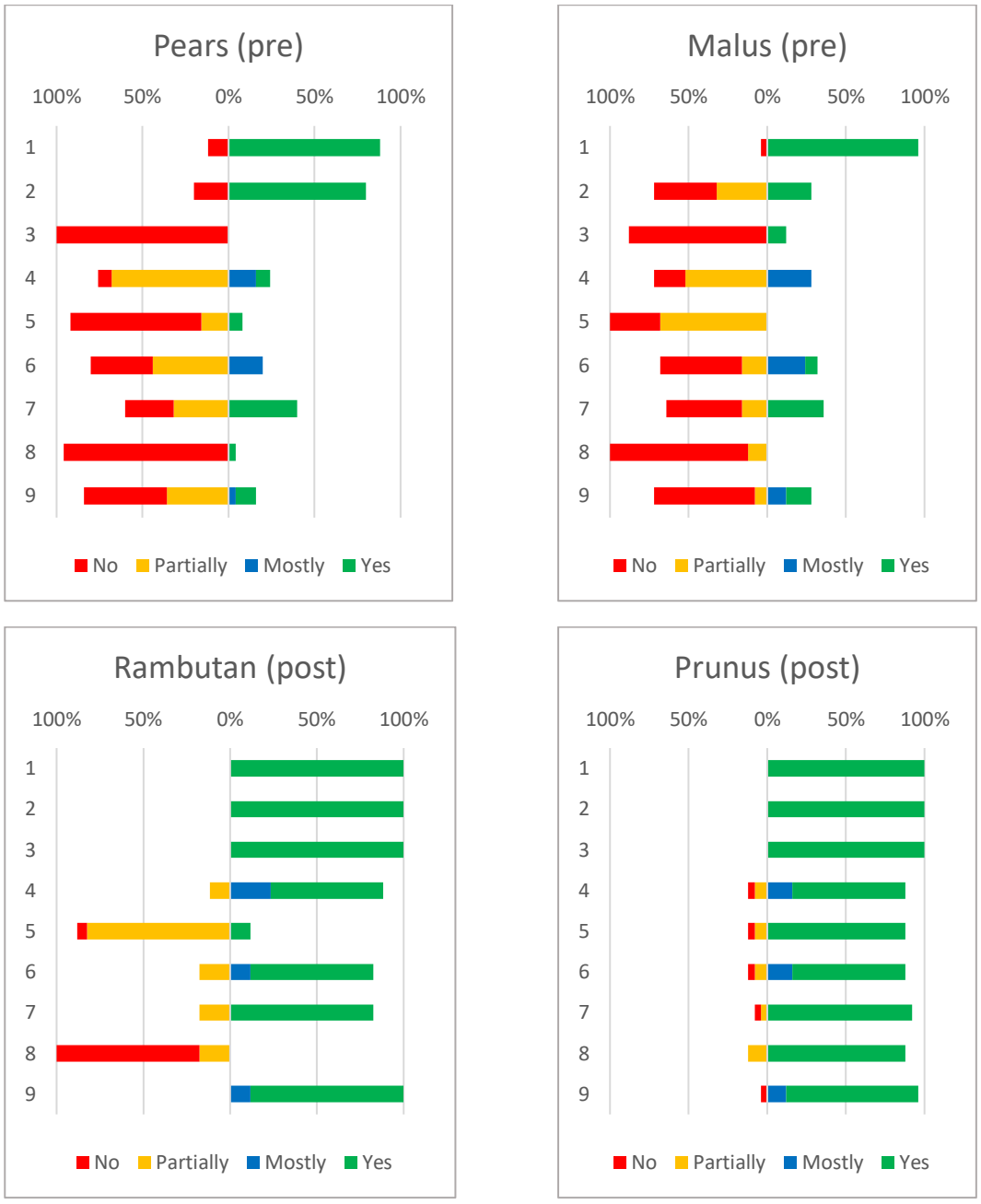
[16] Confidence intervals calculated in Excel using confidence function.

Figure 4-4 shows how pre-CASE CA scores are distributed towards the lower end, and post-CASE CA scores are all above the mid-point. Notably, 17 out of 25 CAs from Prunus for Planting IRA fully instantiated the CASE structure. This was the only report in our sample to have explicitly attempted to structure their analysis with CASE.

The other post-CASE IRA was “Rambutan from Vietnam”. Although this IRA was not explicitly structured with CASE, it seems clear that some elements of CASE were making their way into the drafting.

Another way to get more insight is to look at the distribution of scores on each coding question, comparing the distributions in the pre-CASE reports with those in the post-CASE reports.

Figure 4-5: Distribution of answers to the coding questions.



The four figures above show in more detail the elements of CASE that were lacking in pre-CASE IRAs and present in post-CASE IRAs. As expected, the pre-CASE IRAs failed to instantiate key elements such as

1. 3 - stating the conclusion or contention up front.
2. 5 - ensuring that the reasons are clearly presented as such.
3. 8 - ensuring that evidence presented in favour of the contention is easily identifiable as such.

All three elements might be considered superficial in that they are solely to do with the formatting of the argument. The post-CASE "Rambutan from Vietnam" IRA also did not properly instantiate elements 5 and 8. The Prunus IRA almost always formats and presents arguments according to the CASE Structure.

Why did the Rambutan IRA scored poorly on elements 5 and 8? After all, both these elements are essentially formatting specifications. CASE requires that bullet points should be used in one way only, to indicate that the bulleted item is a piece of evidence in relation to the claim it is nested beneath. The CAs in this IRA use bullet points but all they seem to indicate is a new paragraph. We think the most plausible explanation of this difference is that those involved in drafting made a deliberate (and not unreasonable) choice to use bullet points in line with standard practices at MPI and elsewhere, rather than in distinctive manner required by CASE.

4.3.2 Mixed-effects model

The above descriptive statistics suggest that post-training CAs are associated with higher CASE scores. However, it is possible that this difference could be explained by differences in the type of products under analysis or the types of CAs rated, rather than by the training itself.

To formally test the association between CASE training and the instantiation of CASE in IRAs, we modelled the CASE score for each CA (n = 92) as a function of:

- whether the source IRA was pre- or post-CASE (fixed effect),
- whether the IRA was for plants or produce (fixed effect),
- the type of CA (fixed effect), and
- the IRA identity (random effect).

In other words, we used a linear mixed-effects regression. Of the fixed effect parameters, whether the source IRA was pre- or post-CASE had one of the largest positive effects. Specifically, in the presence of other explanatory variables, post-training CASE scores were estimated to be higher (on average) than pre-training CASE scores by 6.81 points, with a 95% confidence interval of 4.37 to 9.26. The full list of fitted model coefficients is given in Appendix 2 – Study 1 Supplement.

The Analysis of Variance (ANOVA)[17] presented in Table 4-6 below, is similarly suggestive of the significance of the CASE training as an explanatory variable, though the associated p value (0.052) is slightly above the standard threshold of 0.05.

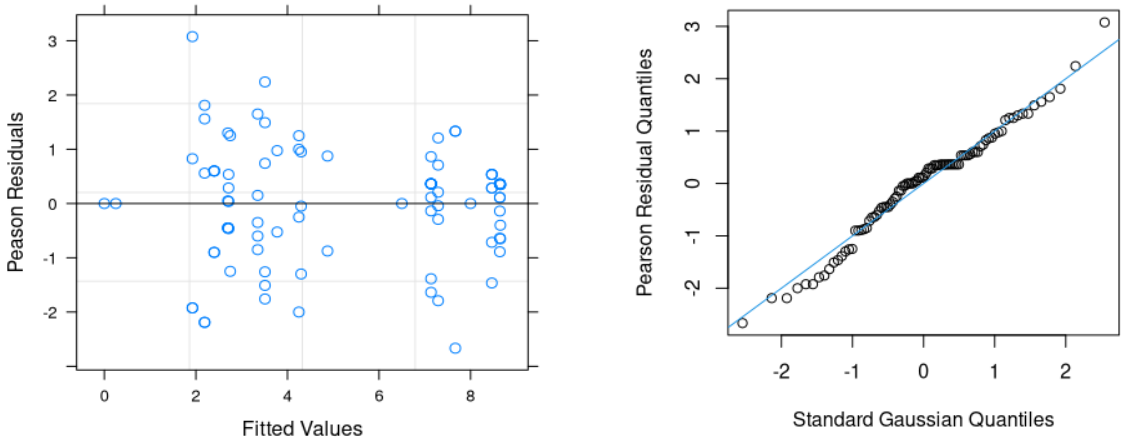
[17] As computed using the anova() function from the lmerTest package in R.

Table 4-6: ANOVA table (using Satterthwaite's method) for the linear mixed-effects regression model of CASE scores in CAs.

Variable	Sum of Squares	Mean Square	Numerator DF	Denominator DF	F value	p value
isPost	5.47	5.47	1	72	3.90	0.052
isProduce	0.35	0.34	1	72	0.25	0.622
typeOfCA	61.50	3.62	17	72	2.60	0.003

The Pearson residuals for the model are shown in Figure 4-6, plotted against the fitted values and standard Gaussian quantiles. There appears to be moderate deviation from Gaussianity, suggesting that the model assumptions do not completely hold, but transformation of the predictor variables is not feasible as all the fixed effect predictors are categorical.

Figure 4-6: Pearson residuals of the model plotted against the fitted values (left), and a Gaussian QQ plot of the same residuals (right).



Goodness-of-fit for mixed effect linear regression models can be quantified using pseudo R^2 statistics. For this fitted model, the marginal R^2, representing the variance explained by the fixed effects, is equal to 0.519. The conditional R^2, representing the variance explained by the full model, is 0.894.[18]

4.4 Discussion

4.4.1 Stronger CASE structure in our samples

Overall, the results show much stronger CASE structure in the post-CASE CAs in our samples.

The CAs in the Prunus for Planting IRA, the only one in our sample that explicitly attempted to apply CASE, instantiated CASE almost perfectly. However CASE was also better instantiated in the other post-CASE IRA, Rambutan from Vietnam. It seems that the CASE initiative was having some effect on drafting of IRAs, even when not being deliberately applied.

It is worth highlighting one difference between pre- and post-CASE CAs in our sample. A key feature of a good argument, whether structured with CASE or not, is a logical link between the conclusion and the detailed information (evidence) provided in its support (see Section 2.3.1.4). That link should

[18] The marginal and conditional R^2 values were calculated using the r.squaredGLMM() function from the MuMIn package in R.

be explicit, so the reader can easily follow the reasoning. We found that this link was often difficult to discern in pre-CASE IRAs. They would present a wealth of evidence without clearly articulating its bearing on the conclusion. This linkage was far more transparent in post-CASE CAs.

We acknowledge that observer bias may have influenced how the CAs were coded. The coder was an employee of the Hunt Laboratory for Intelligence Analysis, the director of which developed CASE and conducted the training. Further, the coder was not blind to whether CAs were pre- or post-CASE. These factors may have caused the coder to see stronger CASE structure in post-CASE IRAs, despite his being aware of the potential for bias and attempting to avoid it. That said, the difference between the pre and post-CASE CAs was dramatic and it is hard to imagine that observer bias accounts for the bulk of that difference.

4.4.2 Generalising to all pre- and post-CASE IRAs

We can generalize the finding in our sample to all reasoning in all pre- and post-CASE Plants division IRAs to the extent that our sample is representative of that larger group.

As discussed in Section 4.2.2, our sampling was purposeful in nature, and so does not constitute a representative sample in a statistical sense. We therefore cannot use standard inferential methods to conclude that the difference found in our sample would also be found in all IRAs, with quantifiable uncertainty.

At a qualitative level, some considerations support representativeness, and some count against it.

One on hand, one of the objectives in the purposeful sampling was ensuring that the sample was as representative as possible, subject to other constraints. To the extent that this was achieved, it was through judgement rather than random sampling. For example, the sample was constructed to ensure that IRAs of the major commodity types were included in both pre- and post-CASE subsamples.

On the other hand, various factors diminish representativeness. These include:

- With only two of the 14 pre-CASE IRAs being included, there is significant room for major differences to arise between the sample and the population due to random variation.
- There was conscious selection bias in our inclusion of the Prunus IRA in the post-CASE sub-sample, on the basis that this was the only IRA in which CASE was deliberately applied, and so offered the greatest prospects for revealing any impact of the CASE initiative. This would skew the results towards showing a greater impact than if all IRAs were used.

Our qualitative assessment is that, taking into account the various factors described above, we can safely conclude that the difference found in our sample reflects a substantial difference between all pre- and post-CASE Plant Division IRAs, though not as strong as our data would suggest.

4.5 Implications

The implications of the results of this study for our research question - To what extent has the CASE initiative improved the clarity and rigour of reasoning in IRAs? - are discussed in Section 7.1.1.

5 Study 2: Do general readers find post-CASE IRAs better-reasoned?

Study 1 found evidence that post-CASE IRAs exhibit stronger CASE structure than pre-CASE IRAs, indicating that post-CASE IRAs do have greater clarity and rigour. However, this difference was identified by an expert in the CASE approach. The "consumers" of IRAs, such as stakeholders and MPI policy makers, generally do not have this particular expertise. Study 2 therefore investigates whether the difference between pre- and post-CASE IRAs is recognizable, by a much wider class of readers, as an improvement in reasoning. It also addresses another problem in Study 1, which was that the expert doing the coding was not independent and may have been biased to find results favourable to CASE.

5.1 Objective

To estimate the extent to which post-CASE IRAs are better reasoned, as judged by general readers.

5.2 Method

5.2.1 Overview

Like Study 1, this study used a retrospective, observational design. We took samples of sections of reasoning from IRAs produced before, and after, the start of the CASE initiative. We recruited participants on a commercial cloud platform to form our sample of general readers. Using a "two alternative forced choice" (2AFC) method, we asked participants to rate sections of reasoning for how well they were "reasoned and communicated." We used statistical analysis to determine whether, and the extent to which, the post-CASE reasoning in our samples were better.

5.2.1.1 The "forced choice" method

2AFC is an experimental paradigm originating in psychophysics (Fechner, Howes, & Boring, 1966) and subsequently adopted in many other branches of psychology (Macmillan & Creelman, 2004). Two stimuli are presented simultaneously to a subject, and the subject must choose between them on some criterion. For example, two images are presented, and the subject must choose which is brighter. The choices made over many such presentations are aggregated to address the research question (e.g., what is the smallest detectable difference in brightness).

The 2AFC paradigm has been used the study of human reasoning. For example, it has been used to test participants' ability to recognise deductive validity (Dube, Rotello, Caren, & Heit, 2010; Trippas, Handley, & Verde, 2014). A recent study at the University of Melbourne tested 2AFC as an alternative to more traditional methods for evaluating the quality of reasoning in short passages comparable in length to the CAs used 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 (ODNI, 2015). Remarkably, participants made their selections extremely quickly (mean selection time of about nine seconds). These results provide some basis for thinking that 2AFC is acceptably valid and efficient means of evaluating the quality of reasoning, and a reasonable choice for this study.

5.2.2 Sampling

This study required us to draw two kinds of samples: a sample of general readers, and samples of reasoning from pre- and post-CASE IRAs to present to those readers.

5.2.2.1 Sampling general readers

We obtained a convenience sample (Battaglia, 2008) of general readers on Amazon Mechanical Turk. This is a crowdsourcing platform where individuals can sign up for paid participation in a wide range of online tasks including human subjects research (Paolacci, Chandler, & Ipeirotis, 2010). Opportunities to partake in such studies are posted to the platform, and interested 'Turkers' can elect to be involved. These Turkers are highly varied in demographics and backgrounds. We recruited 89 Turkers to participate, being paid US$10 per hour.[19]

5.2.2.2 Sampling sections of reasoning (CAs)

For the 2AFC method, we needed samples of reasoning (CAs) from pre- and post-CASE IRAs. To avoid confounding, and to increase statistical power, we wanted these samples to be well matched in relevant features other than being pre- or post. In particular, they should be matched in type in any given forced-choice presentation. For example, if in a given presentation the pre-CASE CA is an "Establishment Assessment" (i.e., assessing the likelihood that a pest will establish in NZ, if it enters), the post-CASE CA should also be an Establishment Assessment.

Of the four IRAs in our sample from Study 1, 2009 Pears from China (pre) and 2019 Prunus for Planting IRA (post) because these two were directly aligned in how they "carved up" the reasoning into CA types. Also, as in Study 1, we used Prunus because it was the only IRA which deliberately applied CASE, and so would likely be most different to a pre-CASE IRA. Table 5-1 summarizes our selection of CAs from these two IRAs. We chose 29 because that was the smallest number such that we could choose at least that many CAs from each type in both IRAs. Where the IRA had more than 29 CA of a type, we randomly selected 29.

Table 5-1: Samples of CAs for presentation in forced choice trials.

Source	Pre or Post	CA Type	Number of CAs
2009 Pears from China IRA	Pre	Establishment Assessment	29
		Assessment of Economic Consequences	29
2019 Prunus for Planting IRA	Post	Establishment Assessment	29
		Assessment of Economic Consequences	29

5.2.3 Procedure

Participants were provided with instructions and training materials prior to the exercise[20]. The training consisted of a short slide deck providing background on the study, guidance on how to make their selections, and worked examples to familiarise themselves with. The guidance outlined criteria against which the CA could be judged. The criteria were presented as questions that should be considered when evaluating the CA, rather than strict rules. The criteria were:

1. Is it clear what the argument is arguing for?
2. Are there clear reasons to believe what is being argued for?
3. Does the argument cite evidence clearly?

[19] Additional selection criteria were minimal. Participants had to be in the US, have previously participated in 100 more similar tasks, and have a 95% completion rate for previous tasks.

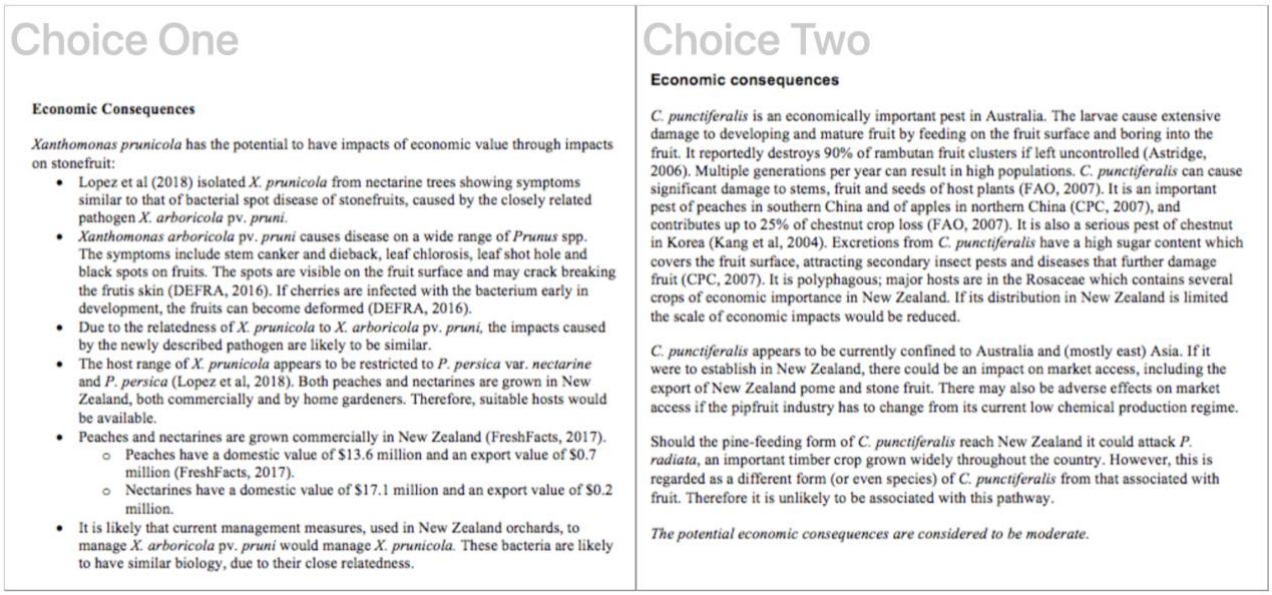
[20] The training module can be found on this project's OSF Repository here: https://osf.io/h4t39/

To help ensure that participants read and absorbed these materials,

- They could not commence the selection tasks until 10 minutes after they first accessed the training.
- They had to correctly answer four multiple-choice questions on the contents of the training materials.

They were then presented with pairs of CAs of the same type. One of the pair was randomly drawn from a pre-CASE report, and the other from a post-CASE report (see Figure 5-1 below). Participants were blind as to whether the CA was produced pre- or post-CASE training. The order of presentation (which was on the left, and which was on the right) was also randomised.

Figure 5-1: Example of what a participant would see in a presentation.



Participants were asked to choose the CA that was "better reasoned and communicated." To help ensure that participants took their time and made sincere choices, we asked them to provide a brief justification for their choice. After they made their selection and provided a justification, they were then presented with another two CAs for comparison. In total, participants made nine selections.

The number of presentations was informed by a power analysis. At a power of 0.8 and alpha of .05 and assuming we want to be able to detect a difference of 5% from a baseline of 50% we needed a minimum of 616 selections.

5.3 Results

5.3.1 Summary statistics

We ended up with a total of 765 usable selections (about 8.5 per participant). This was about 25% more than was required by the power analysis. The post-CASE CA was selected 449 times, or in 59% of selections (95%CI 55%-62%; p<0.0002).

5.3.2 Mixed-effects model

These raw proportions might be confounded by variability within the participants who made the selections. To investigate this we used a binomial (logit) generalised linear mixed-effects regression, modelling an indicator variable for whether the post-CASE CA was selected (n = 765) as a function of:

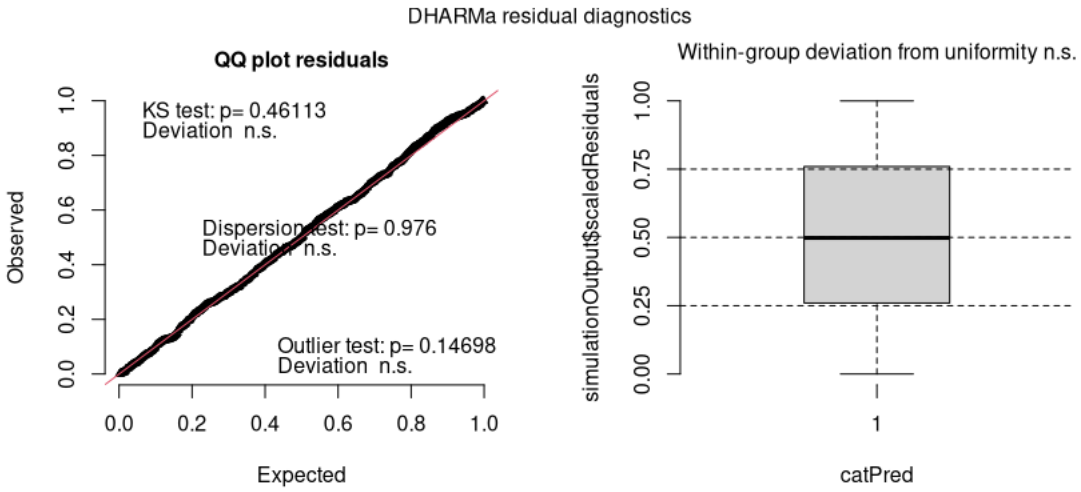
- a constant intercept term (fixed effect), and
- the participant who made the selection (random effect).

Thus controlling for participant effects, the underlying probability that post-CASE CAs were thought to demonstrate superior reasoning was estimated as 0.60 with a 95% confidence interval of (0.54, 0.67). The fact that this estimate (0.60) is so close to the raw proportion (0.59) indicates that the net effect of participant effects is negligible[21].

The effect size (Cohen's *d*) for this difference is 0.31,[22] which is a small-medium effect according to Cohen's widely-used descriptive convention (Cohen, 1988).

To evaluate the fit of the GLMM model described above, we used residual diagnostic plots implemented in the DHARMa package in R. The simulated residual plots in Figure 5-2 indicate that there was no significant deviation from the model assumptions. The (pseudo, conditional) R^2 value for the model was 0.209, indicating that there was a significant amount of variation left unexplained by the single predictor variable.[23]

Figure 5-2: Residual diagnostic plots for the fitted model, as generated by the DHARMa package in R



5.4 Discussion

5.4.1 Better reasoning in our samples

At face value, the results show a small but statistically strong improvement in the quality of reasoning and communication between our pre- and post-CASE samples, as recognised by the participants in our study. However there are various reasons to be cautious about this finding.

First, when tending to select CAs from post-CASE IRAs, participants 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. In further work, this conjecture could be evaluated in a couple of ways. One would be to do a close qualitative analysis of the comments the participants provided when making their selections. Another would be

[21] See Section 10.4 for more details on the mixed-effect model

[22] See Equation 11 in (Stanislaw & Todorov, 1999) for the details of the effect size calculation.

[23] The conditional R^2 value was calculated using the r.squaredGLMM() function from the MuMIn package in R.

do another similar study, but ensure that the pre- and post-CASE CAs were comparable in these superficial features.

Second, participants' selections might have been biased by the nature of the instructions. They were told to select the CA which was "better reasoned and communicated." This succinct wording was chosen for simplicity, but in hindsight it might have been misleading. It seems to treat reasoning and communication as two separate aspects of the CAs. 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 CAs as better communicated in other, non-relevant ways, such as being more "readable," even if the reasoning in those CAs was not as clear. To the extent that this was true, it might mean our data is *understating* the true difference in reasoning quality.

Third, our results might have been biased by the training module. Recall, participants were asked to consider three questions when making their selections:

4. Is it clear what the argument is arguing for?
5. Are there clear reasons to believe what is being argued for?
6. Does the argument cite evidence clearly?

These questions should be answered affirmatively for most good-quality arguments, whether structured with CASE or not. However, these features are all central elements of CASE. It might be argued that these questions biased participants to look for CASE structure – which we know from Study 1 to be more prevalent in post-CASE samples.

It is difficult to estimate the net effect of these factors, because the impact of each is not quantified, and they may to some extent counter-balance. We think a reasonable position is that the post-CASE CAs in our sample were modestly better-reasoned in the eyes of our general readers, but the precise extent is unclear.

5.4.2 Generalising to all pre- and post-CASE IRAs

To what extent the difference identified in our samples be generalised to all pre- versus post-CASE IRAs? As in Study 1, this turns on how representative the IRAs used in this study are of all IRAs. The reasons for caution raised there (Section 4.4.2) also apply here, with the further limitation that in Study 2 we used only two IRAs (one pre, and one post) rather than the four in Study 1. In short, while our purposeful sampling had representativeness as one objective, the residual scope for random variation, and the presence of selection bias (in deliberately choosing the Prunus IRA) mean that our results provide only weak evidence that there would be a similar difference in *all* Plant Division IRAs in the relevant period.

5.4.3 Generalising to other readers

In Study 2 our convenience sample of general readers consisted of "Turkers," or self-selected volunteers on the Amazon Mechanical Turk platform. Turkers are reasonably representative of the population at large – and much more so than the samples of undergraduate students frequently used in social science studies (Paolacci et al., 2010, p. 413). On this basis, and given the very low p value, we infer that members of the public generally would provide responses similar to our convenience sample.

However, the potential readers most relevant to this study come from particular sub-groups of the public, including stakeholders (e.g., importers), or policy makers. We conjecture that such readers are likely to be more educated than the public in general, have more domain knowledge with regard to biosecurity issues, and be more motivated to pay attention to the reasoning in IRAs. We also

conjecture that these factors would make such readers *more* sensitive to differences in the quality of reasoning. Thus the effect found in our sample of general readers is likely to be stronger in the relevant sub-groups.

5.5 Implications

This exploratory study, based on purposeful sampling, provides only weak evidence that post-CASE Plants Division IRAs are better reasoned and communicated in the eyes of general readers than their pre-CASE counterparts. However, it provides stronger evidence that when reasoning has much stronger CASE structure (as in the Prunus IRA, compared with the Pears from China IRA), this difference is in fact better reasoning, as recognised by general readers.

The implications of these results study for our research question - To what extent has the CASE initiative improved the clarity and rigour of reasoning in IRAs? - are discussed in Section 7.1.1.

6 Study 3: Are post-CASE IHS decisions better aligned with risk assessments?

We turn now to the second main research question: to what extent has the improvement in clarity and rigour in IRAs led to better IHS decisions?

Simplistically, to answer this question, we would rate the quality of MPI's IHS decisions, and check whether these ratings are higher in the post-CASE era.

However, it is difficult, in general, to assess the quality of real-world decisions (Keren & de Bruin, 2005). Such decisions typically involve many intuitive judgements, such as what factors to consider and how to weigh them, and there is generally no objective, independent way to rule on the correctness of such judgements. At a deeper level, there is much dispute about what it even means for a decision to be good. For example, are good outcomes more important than good process, or vice versa?

Even if those theoretical problems could be resolved, at a practical level, we clearly did not have the domain knowledge and expertise to rate the overall quality of MPI policy decisions.

For these reasons, we did not attempt to generate, or obtain, ratings of the overall quality of IHS decisions. Rather, we focused on one critical aspect of those decisions. IHS decisions must, by law, take into account the risk assessments presented in IRAs. Hence, the decisions should reflect the levels of risk in those IRAs: the higher the risk, the more stringent (or "severe") the required measures should be, and vice versa. We call this *alignment*.

We cannot expect perfect alignment. IHS decisions must take into account many factors other than the risks as assessed in the IRAs. These other factors might occasionally result in decisions that differ from the ones which would be made if considering only the level of risk. Nevertheless, there should generally be good alignment, alignment "other things being equal."

Thus, in Study 3, we operationalized the main research question as: are the decisions in post-CASE IHS reports *better aligned* with the risk assessments in the corresponding IRAs than the decisions in pre-CASE reports?

This section proceeds first by describing the method used to calculate alignment, before reporting results and finally discussing them. Full details of how each IRA/IHS pair was coded is left to

6.1 Objective

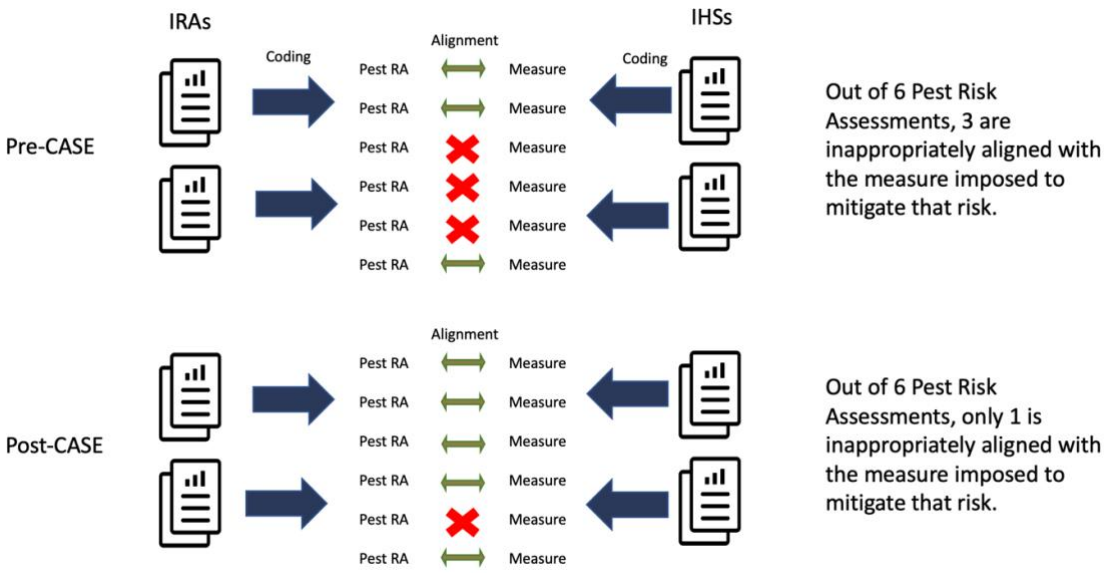
Estimate the extent to which post-CASE IHS decisions were better, in the sense of being better-aligned with the relevant risk assessments.

6.2 Method

6.2.1 Overview

Like the previous studies, Study 3 used a retrospective observational design. We took a purposeful sample of pairs of reports from before, and after, the start of the CASE initiative in 2015. Each pair consisted of an IHS and the relevant IRA. We coded the levels of risk in the IRAs, and the severity of measures specified in the IHS decisions. We then coded each decision for the alignment between the severity of measures it required, and the corresponding level of risk from the IRA. We then used statistical analysis to understand the difference between pre- and post-CASE samples. This approach is illustrated in Figure 6-1.

Figure 6-1: Schematic overview of method for assessing whether CASE adoption has improved decision making. We focus on the alignment between (a) the levels of risk identified for pests in the IRAs, or the need for "extra measures" ("Pest RA"), and (b) the measures required for those pests by decisions in IHS reports ("Measure"). Is alignment better after CASE adoption? The alignment patterns in this figure are entirely made up, for illustrative purposes.



6.2.2 Sampling

6.2.2.1 Sampling of IRA-IHS pairs

We based our sample of IRA-IHS pairs on the purposeful sample of IRAs used in Study 1. That study found a strong difference in the degree to which reasoning in the sampled IRAs exhibited CASE structure (and so, arguably, had greater clarity and rigour of reasoning; see Section 7.1.1). In this study, our immediate objective is to see whether that difference is associated with improved alignment. Thus, our sample of IRA-IHS pairs consisted of the IRAs from Study 1 and the corresponding IHSs, as summarised in Table 6-1.

Table 6-1: Sample of IHS/IRA pairs used in Study 3

IRA	Date	IHS	Date	Pre or Post	Commodity Type
Pears from China	2009	Pears from China	2009	Pre	Fresh Fruit/Vegetables
Malus nursery stock	2012	Importation of Nursery Stock	2020	Pre	Nursery Stock
Rambutan from Vietnam	2016	Rambutan from Vietnam	2016	Post	Fresh Fruit/Vegetables
Prunus for Planting	2019	Prunus for Planting	2019	Post	Nursery Stock

6.2.2.2 Sampling of decisions from within IHS reports

Our sample of decisions essentially consisted of all decisions in the four IHSs listed above. We excluded only those decisions for which there was no corresponding risk assessment in the IRA

paired with the IHS from which the decision was drawn.[24] The number of decisions sampled from each IHS is provided in Table 6-2.

Table 6-2: Decisions constituting our sample.

IHS	Number of decisions
Pears from China	61
Malus Nursery Stock	40
Rambutan from Vietnam	34
Prunus for Planting	31

6.2.3 Procedure

In this section we provide an overview of our coding procedures. These procedures could not be applied in exactly the same way for all four IHSs, given differences between those IHSs. How exactly these procedures were applied in each case is described in Appendix 3 – Study 3 Supplement.

6.2.3.1 Coding risk assessments

The purpose of an IRA is to assess the risk posed by pests associated with an imported commodity. Depending on the commodity, risk is assessed in one of two ways and thus coding the IRAs is also done in one of two ways.

Coding fresh fruit/vegetable IRAs

In fresh fruit/vegetable IRAs, the purpose of an assessment is to determine the level of risk that a pest associated with a fresh fruit/vegetable import would pose to NZ. To determine this, the IRA considers the following factors:

- Entry assessment - Likelihood the pest will be present on a consignment.
- Exposure assessment – Likelihood the pest will be exposed to a suitable NZ host.
- Establishment assessment – Likelihood the pest will establish in NZ once exposed to suitable host.
- Economic consequence assessment – the level of economic impact the pest would have if established.
- Environmental consequence assessment – the level of impact to NZ native species should the pest establish.
- Human health assessment – the potential impact to human health if pest established.

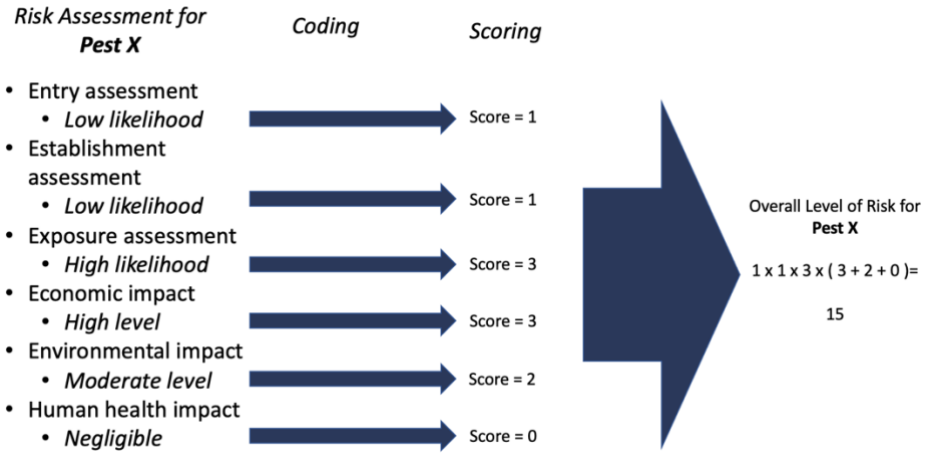
Coding the risk assessments in the fresh fruit/vegetable IRAs was a matter of extracting the likelihood and level of impact from each of the factors and generating an overall risk score for each pest. The IRAs themselves do not provide an overall level of risk; instead, their conclusion is a binary one. That is, from the analysis of risk, the IRA only concludes that the pest is or is not a risk to NZ rather than concluding the overall level of risk is “extremely high” or “moderate”. Importantly, not

[24] Each IHS provides a list of pests of diseases and the required measures that are imposed before a consignment is cleared for import. The list can include pests and diseases that are not assessed in the corresponding IRA. This occurs when a previous IHS has already imposed a risk management measure and no new evidence has come to light that requires the measure be reassessed. Thus the measure from the previous IHS is simply ported into the new one, with no new corresponding risk assessment.

every risk assessment always considers all six of the factors above. If the likelihood of any of the first three factors (entry, exposure, establishment) is considered negligible, then the risk assessment is terminated and the pest is not considered a hazard to NZ.

To generate an overall risk score, we multiplied the three likelihood assessments (entry, exposure and establishment) with each other and with the sum of the impact assessments (economic, environmental, human health).[25] The coding process for each pest is summarised in Figure 6-2.

Figure 6-2: Process for coding risk assessments, with illustrative numbers.



The overall risk score for each pest can then be compared with the associated measure in the IHS to determine if it is aligned.

Coding nursery stock IRAs

We needed a different approach to code the nursery stock IRAs in our sample due their relationship with their corresponding IHS. For both *Malus* and *Prunus* it is a requirement that all nursery stock entering New Zealand be subjected to two growing seasons in post-entry quarantine. Essentially, the risk management decision is the type of test that is conducted on consignment during that time in quarantine. What type of test is applied to the nursery stock depends on the type of pest or disease that is being detected. Thus the difference, for fruit/vegetable decisions, the type of measure depends on the level of risk, for nursery stock it depends on the type of pest.

Although the type of test depends on the type of pest/disease, the number of tests required does depend on the level of risk. This makes sense because if a pest/disease is particularly risky, you want to be as certain as possible that it will be detected, and performing more than one test will provide that assurance.

Nursery stock IRAs are structured with this in mind. In *Malus*, the IRA assesses risk posed by each pest/disease by considering the usual factors (likelihood of entry, establishment and exposure etc.) and based off that assessment a recommendation is made as to the type of test or tests to be conducted. In *Prunus* risk is assessed by examining the likelihood that the pest will go undetected in post-entry quarantine and the damage it would cause if it remained undetected, and based off that assessment, a recommendation is made as to what test or tests must be conducted.

[25] This is similar to the common practice of quantifying risk as the product of likelihood and consequence. However we are developing a more qualitative, coarse-grained risk *score* rather a calculation of risk.

Since the recommendation is based off the assessment of risk, to code the nursery stock IRA we only needed to record the recommendation. For both *Malus* and *Prunus* this was a matter of documenting the test or tests that are recommended to each pest/disease considered in the IRA.

6.2.3.2 Coding decisions

MPI decisions are required measures that mitigate the risk to NZ from pests on imported goods. These decisions are laid out in an IHS document. The IHS specifies which measure(s) need to be applied for all pests associated with an import. The measures themselves vary in stringency, which is to say that one measure is more capable of managing risk than another, though it may be more costly. Take for example, the most stringent measure required when importing pears from China, "reship or destroy and suspend pathway". This measure is extremely stringent because not only will the pears not even enter NZ, but no more pears from that source can be imported.

What measures are available depends on what is being imported. For example, Pears from China require a different set of measures to Prunus plants for planting. So, each IHS required its own coding scheme to reflect the range of measures required to mitigate the risk from specific pests. Coding proceeded by first recording which measure was required for which pest. Each IHS provides a list where this is detailed, so recording was simply a matter of transcribing this list into a database. For fruit/vegetables the next step was to rank the available measures by their stringency, from most capable of managing risk to least. For nursery stock, we recorded what test or tests the IHS requires to be conducted on the consignment is recorded.

6.2.3.3 Coding alignment

Once risk assessments and decisions have been coded, we can code for how aligned the decisions are with the assessments. This is done differently depending on the type of commodity.

Misalignment in fresh fruit/vegetables for import

To recognise alignment, or rather misalignment, and then code for it, we must look at each risk assessment and decision in the context of the others. As we are not subject matter experts, we were not in a position to prescribe which measure should be required for any particular pest. We could, however, analyse a particular pest and its required measure in the context of others. This allowed us to recognise the outliers, where, in context of the measures applied to other pests, one measure did not fit.

To illustrate how outliers are recognised for this commodity type, consider the table below:

Table 6-3: Fictional example of identifying (mis)alignment

Pest	Overall Risk Level	Severity of Measure
B	4	Weak
D	5	Weak
E	6	Weak
G	6	Weak
A	7	Weak
C	8	Severe
H	8	Severe
F	11	Weak

Pest F (shaded) is a potential case of misalignment. It has the highest risk level of all pests and yet only a weak measure is required. On the other hand, pests C and H have lower levels of risk than F and yet require severe measures. To be aligned, pest F would require a severe measure as well and because it doesn't, we can conclude that the risk assessment is misaligned with the decision.

Could pest A be a case of misalignment as well? Its risk level is only slightly lower than C but requires only a weak measure. Pest A is not a case of misalignment because C and H, with a risk level of 8 both require the same measure. The cut off has to be somewhere and in our example, it when the level of risk increases from 7-8. In other words, pest A is a borderline case, not an outlier.

Once a table of the kind just shown has been constructed for the decisions and risk assessments in an IHS/IRA pair, we can calculate the level of misalignment. We do this by counting the number of misalignments and present it as a fraction of total possible misalignments. So, in the example above, there are eight potential cases of misalignment and one case of misalignment and so the overall score would be 1/8.

We can then compare the misalignment scores in IHS/IRA pairs produced pre-CASE adoption and post-CASE adoption. A greater misalignment in pre-CASE report pairs suggests that CASE adoption has improved alignment.

Misalignment in Nursery Stock for Import

To recognise misalignment in nursery stock IRA/IHS pairs, we look to see if the IHS reflects what the IRA suggests. If the IRA suggests that a pest requires extra measures, then the IHS should impose them. If the IHS adopts the measure suggested for a pest in the IRA, then the two are aligned in that regard. Consider the table below as an example.

Table 6-4: Example of how misaligment is recognised in a Nursery Stock IHS

Pest	Suggested Measure in IRA	Measure Imposed in IHS
B	Option 1 + Option 2	Option 1 + Option 2
D	Option 1 + Option 2	Option 1
E	Option 1 + Option 2	Option 1
G	Option 1 + Option 2 + Option 3	Option 1 + Option 2 + Option 3
A	Option 1	Option 1 + 2
C	Option 1 + 2	Option 1 + 2
H	Option 1 + 2	Option 1 + 2
F	Option 1	Option 1

Pest D and E seem to be cases of misalignment. The suggested measure in the IRA does not match what is imposed in the IHS. The IRA has analysed the risk posed by the pest and suggested measures that are capable of mitigating that risk. Despite this, the IHS does not accept the advice of the IRA and imposes measures that may not be capable of managing the risk. Pest A is also a potential case of misalignment. Here the IRA has suggested that option 1 is sufficient to manage the risk posed by the pest, but the IHS imposed an additional measure, option 2.

As with the fruit/vegetables IRA/IHS pairs, the cases of misalignment are summed and presented as a fraction of the total possible cases of misalignment.

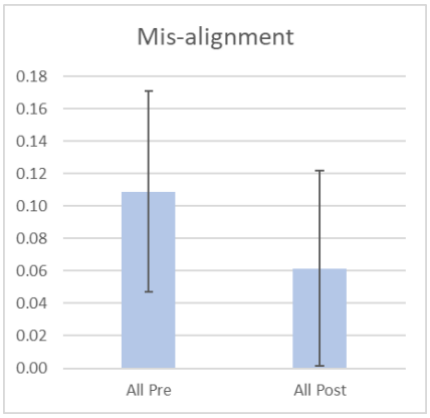
6.3 Results

6.3.1 Summary statistics

Table 6-5: Summary statistics from alignment coding for all IRA/IHS pairs.

Pre or Post	IRA/IHS Pair	Misaligned	% Misaligned with 95% CIs
Pre	Pears from China	4/61	6.6 (0.2, 12.9)
	Malus Nursery Stock	7/40	17.5 (5.2, 29.8)
Post	Rambutan from Vietnam	0/34	0 (0, 0)
	Prunus Plants for Planting	4/31	12.9 (0.4, 25.4)
All Pre		11/101	10.9 (4.7, 17.1)
All Post		4/65	6.2 (0.2, 12.2)

Figure 6-3: Alignment difference (%) for all pre-CASE IRA/IHS pairs, and all post-CASE pairs, in our sample, with 95% CIs.



From these statistics we can see that, overall, the post-CASE pairs have better alignment than the pre-CASE reports, though the difference is small and not statistically significant.

It is interesting to note that the Prunus IRA/IHS pair was the second most misaligned, even though Prunus was the only IRA in which CASE was explicitly applied CASE, and it succeeded in implementing CASE fully. On the other hand, alignment for Prunus was marginally better than for Malus, the corresponding pre-CASE nursery stock IRA. Perhaps something about nursery stock, as compared with fresh fruit/vegetables, leads to increased misalignment.

6.3.2 Binomial regression model

To further assess the significance of the observed increase in alignment following training, we fit a binomial (logit) generalised linear regression model, which models the probability that a given assessment is aligned as a function of:

- whether the assessment was pre- or post-CASE, and
- whether the risk assessment was of type “nursery stock” or “fresh fruit/vegetables”, as the process for evaluating alignment is different in each of those two settings, and we wish to control for this difference.

Fitting such a model[26] produces the following point estimates and 95% confidence intervals for the model parameters. For ease of interpretation, the values have been exponentiated, so that they correspond to odds multipliers.

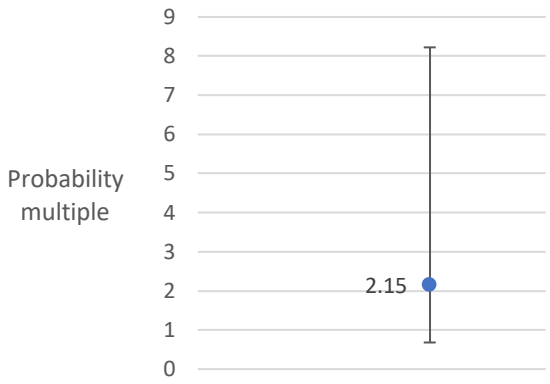
Table 6-6: Binomial logistic regression model of the probability that a given IHS decision is aligned with its corresponding risk analysis.

	Point Estimate	**95% confidence interval**	
		2.5%	**97.5%**
(Intercept)	4.101	2.021	9.202
isPost	2.150	0.681	8.223
isProduce	4.467	1.439	16.867

[26] This model was fit in R using the glm(...) function. The confidence intervals produced use the profile likelihood method.

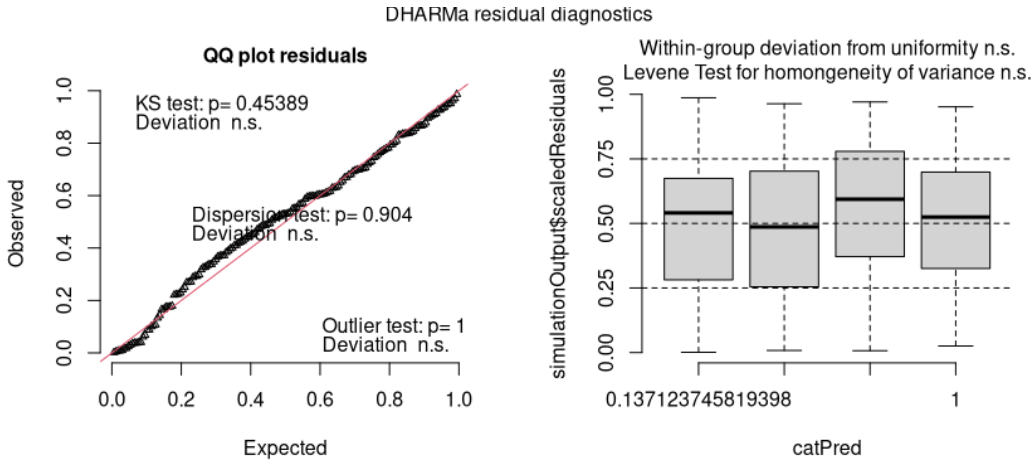
In this model, the odds of alignment post-CASE is 2.15 times the pre-CASE odds. At a significance level of 95%, this is not significantly different from 1.

Figure 6-4: Point estimate of probability (odds) that a post-CASE decision is aligned as a multiple of the probability that a pre-CASE decision is aligned, with 95% confidence interval.



To evaluate the fit of the GLM model described above, we again used residual diagnostic plots implemented in the DHARMa package in R. The simulated residual plots in Figure 6-5 indicate that there was no significant deviation from the model assumptions. The pseudo R^2 value for the model was 0.103, indicating that there was a large amount of variation not explained by the two predictor variables used[27].

Figure 6-5: Residual diagnostic plots for the fitted model, as generated by the DHARMa package in R.



We considered alternate model specifications, which might better make use of the encodings and account for random effects which may be present. However none proved feasible. The small size of the dataset meant that model specifications with additional relevant explanatory variables were singular. Similarly, we considered whether it would be possible to use an ordinal regression framework, modelling the risk measures taken as a function of the assessed risk level. However this wasn't workable because the risk measures are encoded differently for each IRA and cannot be considered to be on the same ordinal scale, and it is not clear how alignment could be quantified in such a framework.

[27] To calculate the R^2 value for the GLM, we used the rsquared() function from the piecewiseSEM R package, with the nagelkerke method.

6.4 Discussion

6.4.1 Alignment in our sample

Results indicate that in general, alignment between risk assessments and corresponding risk management decision is already quite high.

Our results also indicate that alignment is marginally better in post-CASE pairs. The difference is small with an absolute improvement of only 5% (relative improvement 45%), and does not reach the conventional benchmark of statistical significance.

As in Study 1, the data was produced by coding carried out by a single, non-independent expert who was not blind to whether the decisions and assessments were pre- or post-CASE. There is thus potential for lack of reliability and observer bias. However, in this study, the coding was largely mechanical, requiring little judgement, which would substantially mitigate these risks.

6.4.2 Generalising to all pre- and post-CASE IRA/IHS pairs

Given:

- The marginal improvement in alignment found in our sample;
- The relatively small size of our sample of pre-CASE IRA/IHS pairs; and
- Other methodological issues, such as the purposeful nature of our sampling and the potential observer bias

we don't think anything can be inferred about the relative levels of alignment across all pre- and post-CASE Plants Division IRA pairs.

6.5 Implications

The implications of these results study for our research question – To what extent has the improvement reasoning led to better decisions? - are discussed in Section 7.1.2.

7 Conclusion

7.1 Implications

7.1.1 Question 1: To what extent has the CASE initiative improved the clarity and rigour of reasoning in IRAs?

We conducted two exploratory studies aimed at elucidating the extent to which the CASE initiative has improved the clarity and rigour of reasoning in Plant Division IRAs.

Study 1 found that the post-CASE IRAs in our sample had much stronger CASE structure than the pre-CASE IRAs. Taking into account methodological issues, we inferred that there was substantial difference in the level of CASE structure across all pre- and post-CASE Plant Division IRAs, though likely not as sizeable as our data would suggest.

Study 2 found that the post-CASE IRAs in our sample were somewhat better reasoned and communicated in the eyes of independent general readers. Taking into account methodological issues, we regarded this as providing weak evidence of a similar difference across all pre- and post-CASE Plant Division IRAs.

Thus, both studies point towards a change between pre- and post-CASE IRAs, though from different angles and with different strengths. This change appears to be a modest gain in the clarity and rigour of reasoning.[28]

Did the training, practice, and other activities in the CASE initiative *cause* the gain in clarity and rigour, such as it was? This turns on whether factors other than the CASE initiative might have been responsible for that gain. For example, it might have been caused by more intensive review processes which happened to coincide with the CASE initiative.

Our exploratory studies were retrospective and observational in design, using small, purposeful samples. They were unable to control for confounding causal factors, and so were not well-suited to answering causal questions.

We nevertheless deem it likely 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. The nature of the observed change aligns directly with the nature of the efforts to bring about that change. Second, while we can easily imagine confounding factors, such as more intensive review processes, we currently have no positive reason to think that any such factors were actually operating.

7.1.2 Question 2: To what extent has this improvement led to better IHS decisions?

We conducted one exploratory study aimed at elucidating the extent to which the improvement in clarity and rigour led to better decisions in Plant division IHSs.

Study 3 focused on the alignment between policy decisions reported in IHSs and the risk assessments in corresponding IRAs. It found that, in the sampled pre- and post-CASE decisions, post-CASE were better aligned. However this difference was small and statistically non-significant. This would be partly explained by the fact that alignment was already strong in the pre-CASE decisions, so there was not much room for improvement. Taking into account the marginal difference, and other

[28] This is in part a semantic point: increased CASE structure, and increased tendency to be chosen as better reasoned and communicated, *constitute* greater clarity and rigour.

methodological issues, we think our results provide little if any support for an improvement in alignment across all pre- and post-CASE IHS decisions.

Since alignment is only one aspect of quality for IHS decisions, even a substantial improvement in alignment post-CASE would not guarantee that IHS decisions had improved overall. It would mean they had improved in one regard, which would be *prima facie* evidence that they had improved in overall quality.

Suppose we had found strong evidence for substantially better alignment. Could we causally attribute that alignment to the CASE initiative? The issues here are essentially the same as with Question 1. Study 3's design was unable to tease out causal factors, but it would remain plausible that the CASE initiative had a major causal role, until we had positive reason to believe otherwise.

7.1.3 Other implications

As explained in the introduction, the focus of this project was on whether the CASE initiative is helping improve decision making at MPI broadly, and IHS decision making in Plants Division specifically.

Our research did not turn up compelling evidence of such impact, but neither did it turn up compelling evidence against such impact. The issue remains unresolved, due to limitations in scope, scale and design in this exploratory work.

Nevertheless, the project has yielded some insights.

First, the data in all three studies *point towards* a positive impact, even if that impact can't be confidently asserted at this time. That is, the findings are directionally consistent with each other, and with the CASE initiative having helped improved decision making.

Second, we can be confident from Study 1 that at least some post-CASE IRAs exhibit very strong CASE structure, a fact which cannot be plausibly explained except as a result of the CASE initiative. That is, our research has confirmed that MPI *can* (though does not always) produce IRAs which very consistently conform to certain fundamental principles of good reasoning. This is an important finding. An IRA like Prunus establishes a benchmark, for MPI and other similar organisations, with regard to clarity in rigour in the articulation of reasoning in risk assessment work.

Third, Study 3 indicates that plant import health standards decisions are generally well-aligned with the relevant risk assessments. We expect that that this will not be surprising to many at MPI, who would have been able to informally observe it, but we are not aware of alignment having been measured. Our research confirms and quantifies the level of alignment in samples of both pre- and post-CASE IRAs.

Fourth, at a methodological level, our research has increased our understanding of how certain kinds of questions about organisational performance, and to enhance it, can be tackled. Prior to this project, it wasn't at all clear how you could investigate the impact of a program like the CASE initiative on something as important, but impalpable, as the quality of decisions. Our pilot studies were innovative in how the operationalised the problem, and in how they adapted existing research methods to the new context. As pilot exercises, they may be more significant for what they reveal about the challenges involved than for the data they produced. Lessons learned along the way are reflected in Section 7.3, where we discuss potential future research.

7.2 Recommendations

7.2.1 Recommendation 1: Continue the CASE initiative

High-quality biosecurity risk management decisions are critical in balancing the need to protecting the community, economy, and environment from biological threats, on one hand, with avoiding undue restrictions on commerce and on freedom on the other. To strike this balance, risk management decisions must be informed by rigorous risk assessments. These assessments are produced by MPI analysts and managers, and communicated primarily through written reports (IRAs). The analysts and managers generally have strong scientific backgrounds and relevant training and experience. However, only rarely have they had explicit training in articulating and presenting reasoning. This has contributed to risk assessment reports often being not as clear and compelling as they could or should be, hindering their ability to inform decision making.

Given this situation, we recommend that MPI **continue the CASE initiative** – or more broadly, continue to train risk analysts and managers in the use of structured argumentation, and encourage the use of such techniques to improve the clarity and rigour of reasoning in IRAs. The case for CASE can be summarised in five points.

1. In 2015, MPI decided to commence the activities which became the CASE initiative, reflecting managers' assessment at that time that CASE approach was a good fit with their needs. To our knowledge, that assessment is still valid.
2. Our research, the first to attempt to evaluate the initiative, indicates a measurable positive impact of the CASE initiative on reasoning in IRAs, and is consistent with a positive impact on MPI decision making.
3. The CASE initiative is likely to have other benefits. For example, as mentioned in s.3.6, training and practice in argument mapping has been shown to improve generic critical thinking skills. The CASE initiative likely resulted in some such gains, which would plausibly have a broad (albeit hard to measure) impact on aspects of analyst and manager performance beyond IRA drafting.
4. We are not aware of any other strategy by which similar objectives might be achieved more effectively.
5. The CASE initiative could be strengthened, as described below, which would plausibly lead to greater impact.

7.2.2 Recommendation 2: Strengthen the implementation of the initiative

As discussed in Section 2.4, MPI has made a substantial effort to introduce and apply CASE. However, looking back, the initiative evolved in an ad-hoc way, from an initial pilot to a wider roll-out. The level of commitment seems to fluctuate from year to year, depending in part on staff changes. As a result, CASE adoption has been incomplete, and the skills and practices may be faltering.

We therefore recommend that **if MPI does continue with CASE adoption, it should make the implementation more systematic and integrated into MPI training and practices**. Specifically, should consider:

R2.1 Organising more regular training and practice sessions. For example, training might be offered twice yearly. All managers and analysts should have at least a basic level of exposure to the approach.

R2.2 Developing a handbook for the application of CASE principles and techniques in MPI work. Much of this could be adapted from the existing training materials.

R2.3 Verifying knowledge and skills. MPI could make available some form of testing ("certification") to verify that individuals have reached a baseline level of understanding and expertise. Undergoing testing should be voluntary, but a positive result should be recognised as a career-relevant accomplishment.

R2.4 Include CASE-adherence in review. An appropriate level of CASE adherence should be included as another criterion of adequacy in the report review process.

7.2.3 Recommendation 3: Tailor the CASE approach for MPI biosecurity risk analysis

The CASE approach, as introduced to MPI, was a generic framework. It was combined with biosecurity-related examples and exercises, often drawn from IRAs from MPI and DAWE, but the approach itself was neutral with regard to biosecurity generally and the MPI context in particular. Yet it became increasingly clear over time, both to some internal advocates and to the external trainer, that the method itself could be adapted to deliver better outcomes for MPI. However such adaptation would take some effort over and above the activities involved in delivering and implementing the existing ("baseline") method.

We thus recommend that **if MPI does continue with CASE adoption, it should produce a tailored version of the CASE approach adapted to MPI's domain, needs and context**. Specifically, MPI could consider:

R3.1 Develop templates. Develop a "library" of CASE-structured templates for the specific kinds of reasoning patterns which frequently recur in IRA subsections such as entry assessments. These templates could help newcomers to the approach to more easily see how the general CASE structuring principles apply to the kinds of reasoning they deal with; promote consistency in the articulation of reasoning from one analyst to another, and from one IRA to another; and help expedite drafting of IRA subsections.

R3.2 Revise drafting guidelines. The details of how CASE-structured reasoning is expressed in written sections of IRAs should be reviewed and optimised for the MPI context. For example, one senior MPI person voiced a concern that post-CASE CAs, with the main contention stated at the beginning, read as if they are rationalising a pre-determined conclusion rather than expressing a judgement arising from a balanced consideration of all relevant evidence. This kind of concern should be carefully evaluated (e.g., do all/most readers share that sense?), and ways to prevent or mitigate such perceptions evaluated and adopted. This adaptation should be done by CASE experts working closely with experienced MPI analysts, managers, and decision makers. This group should include some who may be sceptical of CASE.

7.2.4 Recommendation 4: Integrate evaluation into the initiative

The CASE initiative proceeded for around five years with little if any systematic evaluation. Our project faced the challenge of evaluating a program retrospectively, in a situation where potentially useful data was lacking (e.g., details of the extent of participation by, or level of CASE expertise of, the risk analysts and managers involved in producing a given IRA).

We thus recommend that **if MPI does continue with CASE adoption, it should integrate evaluation into the initiative itself.** This might occur in three ways:

1. In an ongoing way, systematically gathering data which would be helpful subsequent evaluation activities (e.g., participation data).

2. Conducting some forms of evaluation as part of activities comprising the initiative. For example, participant evaluations of training workshops.
3. Conducting some kinds of research in parallel with, and coordinated with, the initiative. See the next section (7.3.2.1) for further discussion of this opportunity.

7.3 Future Research

There are many opportunities for future research in this area. We divide them into three kinds;

1. Improvements to the studies we conducted.
2. Alternative studies aimed at addressing the same questions
3. New research directions, i.e. research directed at different (albeit related) issues.

7.3.1 Improving Studies 1-3

Studies 1-3 were exploratory in nature. Alongside preliminary findings, they gave us greater insight into the methodological challenges facing this kind of research. There is of course room for improvement and this section describes some of the ways we would improve these studies if we were able to repeat them with more resources.

Study 1

Code a larger sample of pre- and post-CASE IRAs. In Study 1 we coded only four of 18 IRAs produced in the relevant period. Although there were good reasons why we did not code all 18, with more resources we could overcome some of those issues. Also, a future version of Study 1 would have access to a larger pool of post-CASE IRAs.

Modify sampling strategy for greater representativeness. Our purposeful sampling approach, while suitable for this exploratory project, introduced some selection bias. A full version of the study should avoid this by using an alternate strategy, perhaps probabilistic sampling.

Use multiple independent, blinded expert coders. To increase reliability and eliminate potential observer bias, the coding of IRAs for CASE structure should be done by multiple coders who are fully independent of any stakeholders in the success of the CASE initiative, have adequate expertise in structured argumentation, and who are blind to whether the reasoning being coded is pre- or post-CASE.

Use finer-grained data about CASE participation and proficiency. In our studies the critical independent variable was whether an IRA was produced before, or after, the start of the CASE initiative. To better assess the impact of CASE on dependent variables such as the clarity and rigour of reasoning in CAs, it would help to have more detailed information about the extent of participation in the CASE initiative, and levels of proficiency, of individuals involved in drafting IRAs.

Study 2

Use larger sample of pre- and post-CASE IRAs. The reasoning sections (CAs) used in the forced-choice presentations were drawn from just two IRAs, one pre- and one post. For greater generalizability, the sample should be larger and more diverse.

Develop a participant sample more like intended readers. Study 2 aimed to determine whether post-CASE IRAs would appear to general readers to be better reasoned. Our convenience sample of participants resembled the general public, but the actual readers of MPI IRAs belong to special subsets, such as government policy makers. A full version of the study could ensure that study participants more closely resemble such subsets.

Improve materials. The materials we developed and presented to participants could be improved in various ways. One would be to refine the instructions to reduce ambiguity. Another would be to ensure that, in any given forced-choice presentation, the CAs were matched in various ways, such as being of same length, to avoid or mitigate potential confounds.

Analyse subjects' justifications for their choices. We asked participants to provide short (2 sentence) justifications for their choices. This was to help ensure that those choices were thoughtful; we captured those justifications, but did not make any further use of them. In a full version of Study 2, the justifications could be analysed to help determine whether participants' preferences for CAs with stronger CASE structure were driven by substantive differences in quality of reasoning, as opposed to superficial indicators or spurious correlates.

Study 3

This study could be improved in ways very similar to those listed for Study 1.

7.3.2 Alternate studies directed at the same questions

Other types of studies might be used to explore the same (or very similar) questions as addressed in this project. One way to think about this is in terms of high-level study design. Our three studies used retrospective observational designs. Alternatives include experimental designs, or survey- or interview-based designs.

7.3.2.1 Experimental research

In principle, an experiment to investigate the impact of CASE on quality of reasoning in future IRAs could be conducted as follows. It would have two conditions - the experimental or "CASE" condition, and a control condition. In both conditions, one or more teams of analysts and managers would produce one or more IRAs. In the CASE condition, the team(s) would be provided with CASE training, would be certified as having achieved proficiency, and would deliberately apply CASE principles in drafting their IRAs. In the control condition, the team(s) would develop their IRAs without any of these CASE elements. Experimental and control conditions should be similar in all other relevant respects, e.g. they should be developing IRAs on the same commodity types. The resulting IRAS would then be assessed for the clarity and rigour of reasoning, perhaps using methods similar to those deployed in Study 1 and Study 2. An experiment like this would have substantial challenges of its own, but an experimental approach should be better able to isolate the impact of CASE against a background of other factors.

7.3.2.2 Survey or interview-based research

One of the biggest challenges for our project has been obtaining any kind of assessment of the quality of IHS decisions, in order to assess whether those decisions were better post-CASE. A different research strategy would be to tap into the judgement of relevant experts, including the decision makers themselves, and others with suitable experience and expertise. That is, we could use surveys and/or interviews to address issues such as the levels of clarity and rigour in IRAs; whether, in their view, stronger CASE structure constitutes greater clarity and rigour; the quality of IHS decisions; and whether the improved clarity and rigour due to CASE helped improve their decisions. While these methods have their limitations, and probably would be inadequate on their own, they could provide additional insights and potentially address some confounds. For example, in cases where we identified a misalignment between analysis and decision, we could ask the decision maker to explain their rationale. Had they perhaps not found the risk assessment credible? Was the risk credible, but justifiably overridden by other factors?

7.3.3 New research directions

During our research, a variety of new research directions that the current project could serve as a foundation for became apparent.

7.3.3.1 Developing CASE Structured Risk Assessment Templates

As we were coding the CAs in our Study 1 sample, it became apparent that they were using discernible and repeated argument patterns. For example, a significant portion of the Prunus IRA is dedicated to the classification of organisms as hazardous. To make that classification, the organism is evaluated against a set of criteria. The argument pattern is well-understood and can be presented using CASE structure, as in Figure 7-1.

The idea would be to analyse the different types of arguments that are repeated in an IRA and create CASE Structured templates that prompt the analyst to consider the appropriate reasons, evidence and counter-reasons that are required to justify (or undermine) the contention. Exactly what reasons/evidence/counter-reasons are appropriate would be determined by analysing a sufficient number of arguments. These templates could then be used by analysts to streamline development risk assessments.

Figure 7-1: Early sketch of a template for a CASE-structured 'Hazard Identification' argument. Some schematic content has been added just to illustrate how the template might be filled out. Research would be required to develop a suitable set of templates, and to develop a workflow for their use, supported by suitable technology (e.g., Microsoft Word templates or forms).

Contention: Pest X is a hazard.

Reason: It meets sufficient criteria to be considered a hazard.

Sub-reason 1: It meets the first criterion – the pest is absent from New Zealand

- **Evidence:** No record of the organism in databases **Source:** Database X,Y,Z

Bridge 1: If there is no record of an organism in databases then it is reasonable to believe the organism is absent from NZ.

Sub-reason 2: It meets the second criterion – the species has the potential to establish in NZ and harm the economy.

2.1: there are many host species available in NZ for the pest to establish on.

- **Evidence**: Record of pest establishing itself on host species that are in NZ. **Source**: Authors et al.

2.2: The countries where the pest occurs have climates similar to NZ

- **Evidence:** Countries with the pest have a similar CMI (0.7) **Source:** Conservation Biology Institute 2021

2.3: If the pest establishes on a host species, it damages the plant.

- **Evidence:** Infected plants produce less fruit. **Source:** Authors et al.

2.4: Host species are of economic importance to NZ

Evidence: Crop yields from host species are worth $12million a year **Source:** Authors et al.

7.3.3.2 Technology for IRA development

Producing IRAs is a slow and demanding activity, and requires a substantial investment of the organisation's resources. How might it be expedited? The field of artificial intelligence has been making enormous advances in recent years, including in areas such as natural language processing, and argumentation.[29] It is increasingly plausible that tools could help analysts distinguish between reasons and evidence, populate CASE templates, or quickly flesh out a bare-bones argumentative structure into grammatical, well-communicated reasoning. Thus, a major direction for new research is the adaptation of new technologies, particularly AI, into useful tools for biosecurity risk analysts, forming part of their "workbench."[30]

[29] See, for example, the Elicit research assistant (https://elicit.org/) built using OpenAI's GPT-3 language model.

[30] In a current project at the Hunt Laboratory for Intelligence Research, Luke Thorburn is developing a prototype AI-augmented "argument processor" for structured argumentation.

8 Bibliography

Alvarez, C. (2007). *Does Philosophy Improve Reasoning Skills?* (Master's Thesis). University of Melbourne, Melbourne, Australia.

Battaglia, M. (2008). Convenience Sampling. In P. Lavrakas, *Encyclopedia of Survey Research Methods*. 2455 Teller Road, Thousand Oaks California 91320 United States of America: Sage Publications, Inc.

Berry, J. A., Durrant, A., Narouei Khandan, H., Wilson, K., & Philip, B. (2019). *Import Risk Analysis: Prunus plants for planting*. Ministry for Primary Industries, New Zealand.

Biosecurity Act 1993 No 95. , § 23 (1993).

Bloomfield, R., & Bishop, P. (2010). Safety and assurance cases: Past, present and possible future–an Adelard perspective. In *Making Systems Safer* (pp. 51–67). Springer.

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. sage.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Routledge.

Cullen, S., Fan, J., van der Brugge, E., & Elga, A. (2018). Improving analytical reasoning and argument understanding: A quasi-experimental field study of argument visualization. *Npj Science of Learning*, *3*, 1–6.

Dannenberg, A. L. (2016). Effectiveness of health impact assessments: A synthesis of data from five impact evaluation reports. *Preventing Chronic Disease*, *13*.

Dube, C., Rotello, Caren, & Heit, E. (2010). Assessing the belief bias effect with ROCs: It's a response bias effect. *Psychological Review*, *117*, 831–863.

Duffy, J. (2011). Explicit argumentation as a supervisory tool for decision making in child protection cases involving human rights issues. *Practice*, *23*, 31–44.

Ericsson, A., & Pool, R. (2016). *Peak: Secrets from the New Science of Expertise*. Houghton Mifflin Harcourt.

Fechner, G. T., Howes, D. H., & Boring, E. G. (1966). *Elements of psychophysics* (Vol. 1). Holt, Rinehart and Winston New York.

Franqueira, V. N. L., & Horsman, G. (2020). Towards Sound Forensic Arguments: Structured Argumentation Applied to Digital Forensics Practice. *Forensic Science International: Digital Investigation*, *32*, 300923.

Gagnon, M.-P., Desmartis, M., Poder, T., & Witteman, W. (2014). Effects and repercussions of local/hospital-based health technology assessment (HTA): A systematic review. *Systematic Reviews*, *3*, 1–14.

Haley, C., Laney, R., Moffett, J., & Nuseibeh, B. (2008). Security requirements engineering: A framework for representation and analysis. *IEEE Transactions on Software Engineering*, *34*, 133–153.

Kelly, T. (2004). A systematic approach to safety case management. *SAE Transactions*, 257–266.

Keren, G., & de Bruin, W. B. (2005). On the Assessment of Decision Quality: Considerations Regarding Utility, Conflict and Accountability. In D. Hardman & L. Macchi (Eds.), *Thinking: Psychological Perspectives on Reasoning, Judgment and Decision Making* (pp. 347–363). Chichester, UK: John Wiley & Sons, Ltd.

Kirschner, P., Buckingham Shum, S., & Carr, C. (2002). *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*. London U.K.: Springer.

Kneupper, C. W. (1978). Teaching Argument: An Introduction to the Toulmin Model. *College Composition and Communication*, *29*, 237–241. JSTOR.

Langer, L., Tripney, J., & Gough, D. (2016). *The science of using science: Researching the use of research evidence in decision-making*.

Lavis, J., Davies, H., Oxman, A., Denis, J.-L., Golden-Biddle, K., & Ferlie, E. (2005). Towards systematic reviews that inform health care management and policy-making. *Journal of Health Services Research & Policy*, *10*, 35–48.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology Press.

MAF Biosecurity NZ. (2010). *Import Health Standard Commodity Sub-class: Fresh Fruit/Vegetables Pyrus bretschneideri, Pyrus sp. Nr. Communis and Pyrus pyrifolia from the People's Republic of China*.

MAF Biosecurity NZ. (2016). *Risk Management Proposal: Fresh Rambutan for Consumption*.

MAF Biosecurity NZ. (2019). *Risk Management Proposal: Prunus Plants for Planting*.

MAF Biosecurity NZ. (2020). *Import Health Standard: Prunus Plants for Planting*.

Mahtani, K., Spencer, E. A., Brassey, J., & Heneghan, C. (2018). Catalogue of bias: Observer bias. *BMJ Evidence-Based Medicine*, *23*, 23.

Ministry for Primary Industries New Zealand. (2020a). Import health standards.

Ministry for Primary Industries New Zealand. (2020b). Import risk analysis.

Minto, B. (2009). *The pyramid principle: Logic in writing and thinking*. Pearson Education.

Mitton, C., Adair, C. E., Mckenzie, E., Patten, S. B., & Perry, B. W. (2007). Knowledge Transfer and Exchange: Review and Synthesis of the Literature. *The Milbank Quarterly*, *85*, 729–768.

Narrative Literature Review. (2017). In M. Allen, *The SAGE Encyclopedia of Communication Research Methods*. Thousand Oaks California: SAGE Publications, Inc.

ODNI. (2015). *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.

Okada, A., Buckingham Shum, S., & Sherborne, T. (2008). *Knowledge Cartography: Software Tools and Mapping Techniques*. London: Springer.

Orton, L., Lloyd-Williams, F., Taylor-Robinson, D., O'Flaherty, M., & Capewell, S. (2011). The Use of Research Evidence in Public Health Decision Making Processes: Systematic Review. *PLOS ONE*, *6*, e21704.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision Making*, *5*, 411–419.

Patton, M. Q. (2015). *Qualitative research & evaluation methods: Integrating theory and practice.* (Fourth edition.). SAGE Publications, Inc.

Pineda, P. (2010). Evaluation of training in organisations: A proposal for an integrated model. *Journal of European Industrial Training*, *34*, 673–693.

Poder, T. G., Bellemare, C. A., Bédard, S. K., Fisette, J.-F., & Dagenais, P. (2018). 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*, 393–399.

Rider, Y., & Thomason, N. (2008). Cognitive and Pedagogical Benefits of Argument Mapping: L.A.M.P. Guides the Way to Better Thinking. In A. Okada, S. Buckingham Shum, & T. Sherborne (Eds.), *Knowledge Cartography: Software Tools and Mapping Techniques* (pp. 113–130). Springer.

Scriven, M. (1976). *Reasoning*. New York: McGraw-Hill.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*, 137–149.

Thomason, N. (2014). *Critical Thinking and Argument Mapping—Report on Intelligence Advanced Research Projects Activity contract IARPA-BAA-10-08*. University of Melbourne.

Toulmin, S. E. (2003). *The Uses of Argument* (2nd ed.). Cambridge: Cambridge University Press.

Trippas, D., Handley, S. J., & Verde, M. F. (2014). Fluency and belief bias in deductive reasoning: New indices for old effects. *Frontiers in Psychology*, *5*, 631.

Twardy, C. (2004). Argument maps improve critical thinking. *Teaching Philosophy*, *27*, 95–116.

Tyson, J., Rainey, S., Breach, J., & Toy, S. (2009). *Import Risk Analysis: Pears (Pyrus bretschneideri, Pyrus pyrifolia, and Pyrus sp. Nr. Communis) fresh fruit from China*. Ministry for Primary Industries, New Zealand.

van Gelder, T. J. (2002). Enhancing Deliberation Through Computer-Supported Argument Visualization. In P. Kirschner, S. Buckingham Shum, & C. Carr (Eds.), *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making* (pp. 97–115). London: Springer-Verlag.

van Gelder, T. J. (2016, April). Mapping an argument: Dispelling the curse of knowledge. *Decision Point*, (95). Retrieved from http://decision-point.com.au/article/mapping-an-argument/

van Gelder, T. J. (2019). *Making the CASE: Argument mapping for better reasoning and communication [training material]*.

van Gelder, T. J. (n.d.). *Argument Mapping Short Course*. Retrieved from www.vangeldermonk.com

van Gelder, T. J., Bissett, M., & Cumming, G. (2004). Cultivating Expertise in Informal Reasoning. *Canadian Journal of Experimental Psychology*, *58*, 142–152.

Walton, D., Reed, C., & Macagno, F. (2008). *Argumentation Schemes*. Cambridge: Cambridge University Press. Cambridge Core.

Zechmeister, I., & Schumacher, I. (2012). The impact of health technology assessment reports on decision making in Austria. *International Journal of Technology Assessment in Health Care*, *28*, 77–84.

Appendix 1 – Background Supplement

Here we provide detailed material that will help the reader understand how CASE modifies the structure and presentation of an argument in the biosecurity context. Also reported here are the all the factors that must be considered when developing an IHS as mandated by the Biosecurity Act 1993.

9.1 Example of CASE-structured reasoning

This "before and after" example was provided to CASE workshop participants as part of their training materials.

G.uvicola – pre-CASE

The text below was taken from a draft DAWR document, produced before any CASE training.

The likelihood that *G. uvicola* will arrive in Western Australia with the importation of table grapes from India is: **High**.

The following information provides supporting evidence for this assessment.

- *Greeneria uvicola* has been recorded in Andhra Pradesh, Bihar (Reddy and Reddy 1983) and Karnataka (Ullasa and Rawal 1986). Andhra Pradesh and Karnataka are commercial grape production areas expected to export grapes to Australia (DPP 2009).
- *Greeneria uvicola* infects grape clusters (McGrew 1988). On young berries, symptoms first develop as brown lesions (Milholland 1991) or flecks (Kummuang *et al.* 1996b). Severe infection can cause blight on young berries and pedicels which causes young berries to shrivel and drop (McGrew 1988; Kummuang *et al.* 1996b; Momol *et al.* 2007).
- On maturing berries, the fungus causes brownish, water-soaked lesions, with concentric rings of spore bodies, which rapidly spread and eventually cover the entire berry (Momol *et al.* 2007; Ellis 2008; Taylor 2012). Black, raised acervuli form on the decaying fruit which can cause the epidermis and cuticle to rupture (McGrew 1988; Momol *et al.* 2007). Some infected berries soften and detach easily from the bunch, particularly in wet weather, whilst others continue to dry and shrivel (Ullasa and Rawal 1986; McGrew 1988; Momol *et al.* 2007; Taylor 2012). Grape bunches with several berries missing, or with several shrivelled berries, are likely to be discarded at harvesting or packing processes.
- Symptoms of infection are easily recognised on the berries and are reported to develop on healthy berries one week after contact with fungal spores and in less time on damaged fruit (Castillo-Pando *et al.* 1999; Ellis 2008). However, one study which pinned bitter-rotted berries onto healthy bunches did not result in infection of adjacent non-wounded berries (Ridings and Clayton 1970). Infected grape berries/bunches showing obvious symptoms are likely to be removed from the export pathway during harvesting or packing processes. It has also been reported that grapes inoculated with *G. uvicola* from bloom to two weeks before harvest did not show symptoms until just close to harvest (Longland and Sutton 2008). Some authors report that *G. uvicola* invades pedicels of grapes in the spring (shortly after flowering) but remains latent until the berry reaches maturity (McGrew 1988; Momol *et al.* 2007). The fungus then invades the berries, where conidia are produced within four days (McGrew 1988). Kummuang *et al.* (1996b) also reported

that *G. uvicola* was isolated from symptomless berries, especially those late in the growing season. Infected grape bunches without or with only mild symptoms at harvest may escape detection and be picked and packed for export.

- The fungus can invade any injured tissue of *Vitis* spp. plants (McGrew 1988). Injury to mature, healthy berries due to bird and insect damage or cracking of berries due to rain can allow conidial infection and lead to rapid spread of the disease (McGrew 1988; Momol *et al.* 2007). Damaged grape berries/bunches are likely to be removed from the export pathway during harvesting or packing processes.
- The varieties known to be naturally infected in India are Anab-e-Shahi, Angur Kalan, Black Champa, Gulabi, Jaos Beli, Kali Sahabi, Khandari, Pandri Sahebi, Selection 94, Thompson Seedless and Taifi Rosovi (Reddy and Reddy 1983). Some of these varieties are likely to be exported to Australia (DPP 2007; DPP 2009).
- Measures used to control *G. uvicola* in India include pruning of infected canes (NHB 2011).
- Bitter rot symptoms develop quickly on mature berries. It could be expected that any berries with latent infection that were picked and packed for export via sea freight would show symptoms by the time they arrive in Western Australia. Grape bunches showing symptoms would be detected during routine inspection on arrival. However, grapes are usually stored at low temperatures to prolong shelf life. Information on the time required for symptoms to develop under cold storage conditions could not be found, but it is likely that symptoms will develop more slowly under low temperatures. Grapes via air freight may show no or mild symptoms at the time they arrive in Western Australia. Grape bunches without symptoms, or with only minor symptoms, may not be detected at routine inspection on arrival.

The possibility for some late infected berries to show no or mild symptoms and the uncertainty about the development of symptoms at low temperatures support a likelihood estimate for importation of “high”.

G.uvicola – CASE Format

The text below is almost the same content and wording, but has CASE structure applied.

The likelihood that *G. uvicola* will arrive in Western Australia with the importation of table grapes from India is high.

This is because it is highly likely that some grapes harvested in India will be infected, it is highly likely that some of those will be packed for export, and it is highly likely that some of those infected grapes will arrive in WA. Taking these points in turn:

It is highly likely some harvested grapes will be infected by G.Uvicola.

This is because G.uvicola is present in Indian grape-growing regions

- It has been recorded in Andhra Pradesh (Reddy and Reddy 1983) and Karnatka (Ullasa and Rawal 1986).
- Andhra Pradesh and Karnataka are commercial grape production areas expected to export grapes to Australia (DPP 2009)

and it affects grape varieties likely to be harvested for export to Australia:

- The varieties known to be naturally infected in India are Anab-e-Shahi, Angur Kalan, Black Champa, Gulabi, Jaos Beli, Kali Sahabi, Khandari, Pandri Sahebi, Selection 94, Thompson Seedless and Taifi Rosovi (Reddy and Reddy 1983)
- and some of these varieties are likely to be imported into Australia (DPP 2007; DPP 2009).

And, it is highly likely that some infected grapes will be packed for export.

It is true that *G. uvicola* generally causes obvious symptoms in grapes:

- On young berries, symptoms first develop as brown lesions (Milholland 1991) or flecks (Kummuang et al. 1996b)
- Severe infection can cause blight on young berries and pedicels which causes young berries to shrivel and drop (McGrew 1988; Kummuang et al. 1996b; Momol et al. 2007).
- On maturing berries, the fungus causes brownish, water-soaked lesions, with concentric rings of spore bodies, which rapidly spread and eventually cover the entire berry (Momol et al. 2007; Ellis 2008; Taylor 2012)
- Black, raised acervuli form on the decaying fruit which can cause the epidermis and cuticle to rupture (McGrew 1988; Momol et al. 2007)
- Some infected berries soften and detach easily from the bunch, particularly in wet weather, whilst others continue to dry and shrivel (Ullasa and Rawal 1986; McGrew 1988; Momol et al. 2007; Taylor 2012)

and grape bunches showing obvious symptoms are likely to be removed during harvesting or packing processes.

However the main reason that some infected grapes are likely be packed for export is that G. uvicola may be present in symptomless berries:

- Symptoms of infection may take up to one week to appear on healthy berries (Castillo-Pando et al 1999; Ellis 2008)
- Grapes inoculated with G. uvicola from bloom to two weeks before harvest did not show symptoms until just close to harvest (Longland and Sutton 2008).
- G. uvicola invades pedicels of grapes in the spring (shortly after flowering) but remains latent until the berry reaches maturity (McGrew 1998; Momol et al. 2007)

- G. uvicola was isolated from symptomless berries, especially those late in the growing season. (Kummuang et al. 1996b)

and grape bunches showing no visible symptoms are unlikely to be removed during harvesting or packing processes.

And, it is highly likely that some of those infected grapes will arrive in WA.

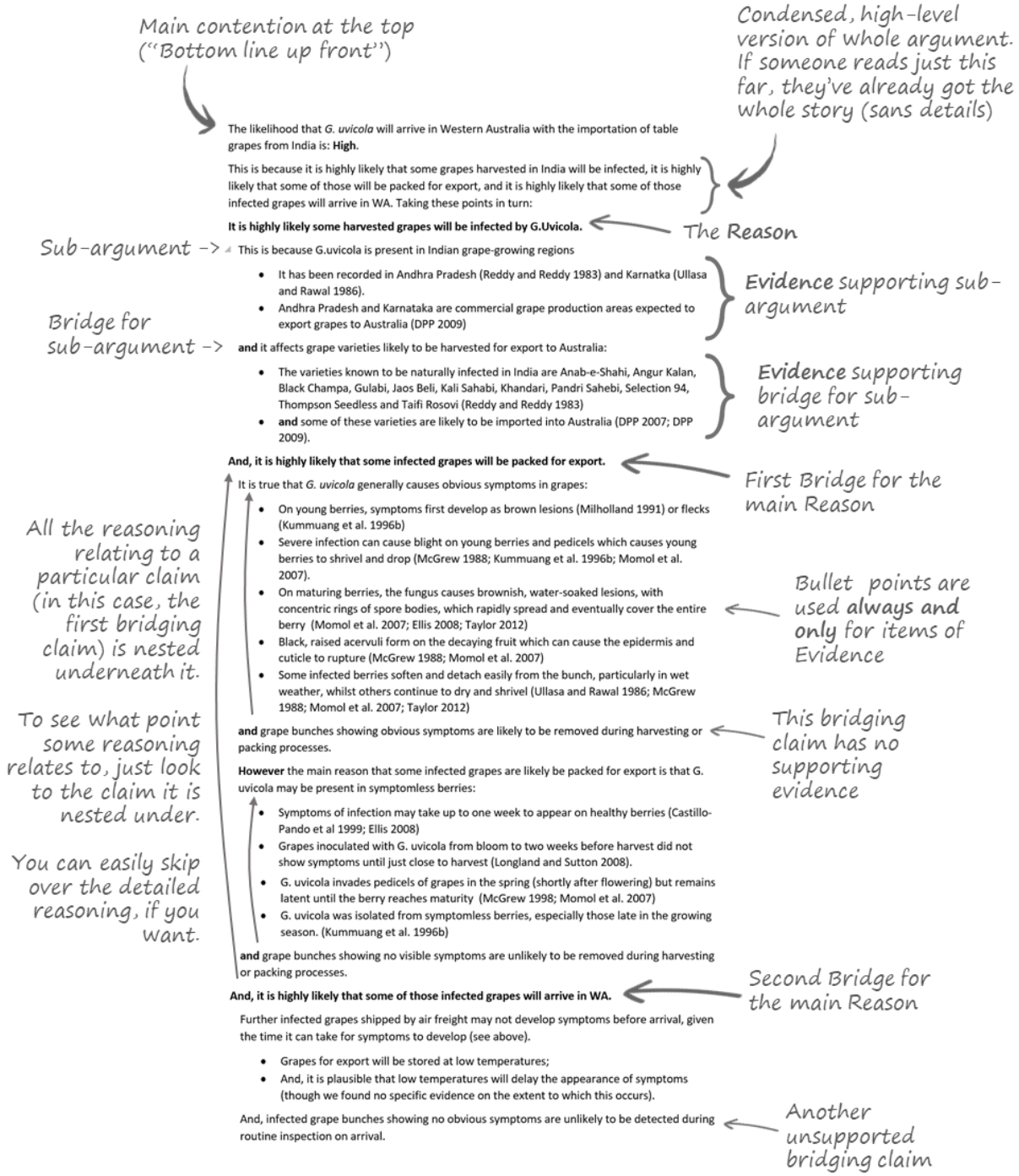
Grapes for export will be stored at low temperatures, and it is plausible that low temperatures will delay the appearance of symptoms (though we found no specific evidence on the extent to which this occurs).

Further infected grapes shipped by air freight may not develop symptoms before arrival, given the time it can take for symptoms to develop (see above).

Infected grape bunches showing no obvious symptoms are unlikely to be detected during routine inspection on arrival.

G.Uvicola – Annotated

This shows the CASE-formatted version, annotated to explain how it exhibits CASE structure.



9.2 Full list of legislated considerations

"In the course of developing the version of the standard for recommendation to the Director-General, the officer—

(a) must have regard to the matters raised by the persons consulted; and

(b) must have regard to the following matters in relation to craft of the class or description proposed for coverage by the standard:

(i) the likelihood that the craft will import organisms into New Zealand territory:

(ii) the nature of the organisms that the craft may import into New Zealand territory:

(iii) the possible effect on human health, the New Zealand environment, and the New Zealand economy of the organisms that the craft may import into New Zealand territory:

(iv) New Zealand's obligations under international agreements; and

(c) must have regard to the following matters in relation to craft of the class or description proposed for coverage by the standard and the requirements proposed for inclusion in the standard:

(i) the extent to which the requirements reduce or manage the likelihood of adverse effects from organisms that may be imported in or on the craft:

(ii) the extent to which the requirements reduce or manage the impacts of adverse effects from organisms that may be imported in or on the craft; and

(d) may have regard to the following matters in relation to craft of the class or description proposed for coverage by the standard and the requirements proposed for inclusion in the standard:

(i) the direct cost of the requirements on owners or operators, or the persons in charge, of craft:

(ii) the direct cost of the requirements on the Crown:

(iii) other economic factors involved in implementing the requirements:

(iv) technical and operational factors involved in implementing the requirements; and

(e) may have regard to any other matters that the officer considers relevant to achieving the purpose of this Part" (Biosecurity Act 1993 No 95, 1995).

Appendix 2 – Study 1 Supplement

Detail on the coding process as well as results from coding the 4 IRAs can be found here.

10.1 Full list of coding questions, possible answers and corresponding scores.

	Question	Possible Answers	Score
1.	Is there a main contention?	Yes	1
		No	0
2.	Is the main contention easily identifiable as such?	Yes	1
		Partially	0.5
		No	0
3.	Is the main contention positioned at the top?	Yes	1
		No	0
4.	Is there a high-level argument structure?	Yes	1
		Mostly	0.75
		Partially	0.25
		No	0
5.	Is the high-level argument structure easily identifiable as such?	Yes	1
		Partially	0.5
		No	0
6.	If there are high-level arguments, is it clear which items of evidence, if any, are intended to support the arguments?	Yes	1
		Mostly	0.75
		Partially	0.25
		No	0
7.	Are the items of evidence clearly relevant to the reasons?	Yes	1
		Partially	0.5
		No	0
8.	Are items of evidence clearly presented as such?	Yes	1
		Partially	0.5
		No	0
9.	Is each item of evidence appropriately sourced?	Yes	1
		Mostly	0.75
		Partially	0.25
		No	0

10.2 CA types coded and their frequency

IRA	CA Type	Number of CA Coded	Frequency in IRA
2009 Pears from China	Bagging of Fruit	1	32
	Cold Treatment	1	25
	Economic Consequences	4	31
	Entry Assessment	4	61
	Environmental Consequences	5	30
	Establishment Assessment	2	34
	Exposure Assessment	5	34
	Pest Free Areas	2	32
	Uncertainty	1	7
	Total	25	
2012 Malus Nursery Stock	Entry assessment	3	31
	Exposure + Establishment	5	31
	Spread assessment	4	31
	Economic Consequences	6	29
	Environmental Consequences	2	30
	Human health consequences	5	27
	Total	25	
2016 Rambutan from Vietnam	Impacts	6	6
	Introduction Assessment	11	11
	Total	17	
2019 Prunus for Planting	Conclusion	3	9
	Risk assessment against criteria for requiring additional measures	5	21
	Hazard identification: commodity association	7	22
	Hazard identification: quarantine pest status	9	38
	Conclusion summary	1	39
	Total	25	

10.3 CASE Adoption Full Coding Results

Source	Type	Overall Score
2009 Pears from China	Environmental Consequences	0
	Establishment Assessment	5.75
	Environmental Consequences	4
	Entry Assessment	2.25
	Entry Assessment	3.25
	Bagging of Fruit	0.25
	Entry Assessment	3
	Environmental Consequences	2.75
	Exposure Assessment	3.5
	Pest Free Areas	4
	Environmental Consequences	3.75
	Pest Free Areas	1.5
	Entry Assessment	2.75
	Exposure Assessment	2.75
	Exposure Assessment	3
	Exposure Assessment	2.5
	Economic Consequences	2.75
	Economic Consequences	0
	Uncertainty	6.5
	Economic Consequences	0
	Exposure Assessment	5
	Cold Treatment	0
	Establishment Assessment	4
	Environmental Consequences	0
	Economic Consequences	5

2012 Malus Nursery Stock	Exposure + Establishment	2.25
	Exposure + Establishment	2.25
	Economic Consequences	1.75
	Human Health Consequences	3
	Spread Assessment	4
	Spread Assessment	5.5
	Human Health Consequences	3
	Economic Consequences	2.25
	Spread Assessment	2.25
	Entry Assessment	3
	Exposure + Establishment	2.25
	Entry Assessment	5.25
	Exposure + Establishment	4
	Spread Assessment	5.25
	Human Health Consequences	1.5
	Economic Consequences	2
	Entry Assessment	4.25
	Environmental Consequences	4.75
	Human Health Consequences	1.5
	Economic Consequences	5.75
	Environmental Consequences	3.25
	Economic Consequences	4.25
	Exposure + Establishment	2.75
	Human Health Consequences	3
	Economic Consequences	5
2016 IRA - Rambutan from Vietnam (note that here CA types are simplified. In Rambutan IRA, CA titles are sentences e.g. “pest X is likely to have a moderate impact on NZ economy”	Impacts	8.5
	Impacts	7
	Impacts	8
	Introduction Assessment	7.5
	Introduction Assessment	7.5
	Introduction Assessment	5.75
	Introduction Assessment	5.5
	Introduction Assessment	7.5
	Impacts	7.5
	Impacts	7.25
	introduction Assessment	7

	Impacts	5.5
	Introduction Assessment	7.5
	Introduction Assessment	7.25
	Introduction Assessment	8
	Introduction Assessment	7.5
	Introduction Assessment	7.5
2019 IRA – Prunus for Planting	Hazard identification: commodity association	8.75
	Hazard identification: commodity association	7.75
	Hazard identification: commodity association	8.75
	Hazard identification: commodity association	9
	Hazard identification: quarantine pest status	9
	Hazard identification: quarantine pest status	8
	Hazard identification: quarantine pest status	9
	Risk assessment against criteria for requiring additional measures	8
	Hazard identification: quarantine pest status	7.75
	Risk assessment against criteria for requiring additional measures	8.25
	Hazard identification: quarantine pest status	9
	Hazard identification: quarantine pest status	8.75
	Hazard identification: quarantine pest status	8.75
	Conclusion summary	8
	Conclusion	9
	Hazard identification: quarantine pest status	8.5
	Hazard identification: commodity association	9
	Risk assessment against criteria for requiring additional measures	9
	Risk assessment against criteria for requiring additional measures	9
	Hazard identification: commodity association	7
	Risk assessment against criteria for requiring additional measures	9
	Conclusion	5
	Hazard identification: commodity association	9
	Conclusion	9
	Hazard identification: quarantine pest status	9

10.4 Mixed-effects model details

Variable	Estimate	t value	p value	95% CI
(Intercept)	1.835	0.712	0.479	(-0.421, 4.091)
isPost	6.815	1.974	0.052	(4.375, 9.255)
isProduce	-1.585	-0.495	0.621	(-2.468, -0.702)
typeOfCA (Cold Treatment)	-0.250	-0.149	0.882	(-3.186, 2.686)
typeOfCA (Conclusion)	-0.983	-1.136	0.260	(-2.499, 0.533)
typeOfCA (Conclusion summary)	-0.650	-0.501	0.618	(-2.924, 1.624)
typeOfCA (Economic Consequences)	1.674	1.309	0.195	(-0.567, 3.915)
typeOfCA (Entry Assessment)	2.464	1.917	0.059	(0.212, 4.715)
typeOfCA (Environmental Consequences)	1.940	1.522	0.132	(-0.294, 4.174)
typeOfCA (Establishment Assessment)	4.625	3.187	0.002	(2.082, 7.168)
typeOfCA (Exposure + Establishment)	0.865	0.621	0.536	(-1.574, 3.304)
typeOfCA (Exposure Assessment)	3.100	2.388	0.020	(0.826, 5.374)
typeOfCA (Hazard identification: commodity association)	-0.186	-0.268	0.790	(-1.401, 1.030)
typeOfCA (Hazard identification: quarantine pest status)	-0.011	-0.017	0.987	(-1.169, 1.147)
typeOfCA (Human Health Consequences)	0.565	0.406	0.686	(-1.874, 3.004)
typeOfCA (Impacts)	7.042	2.066	0.042	(4.799, 9.284)
typeOfCA (Introduction Assessment)	6.886	2.030	0.046	(4.718, 9.055)
typeOfCA (Pest Free Areas)	2.500	1.723	0.089	(-0.043, 5.043)
typeOfCA (Spread Assessment)	2.415	1.704	0.092	(-0.068, 4.898)
typeOfCA (Uncertainty)	6.250	3.730	< 0.001	(3.314, 9.186)

Appendix 3 – Study 3 Supplement

This appendix provides a detailed look at the processes undertaken to code the IRA/IHS pairs in our sample. The coding process to extract the necessary components to calculate alignment is slightly different for each pair and these are reported on in full in this section.

11.1 Coding the Pears from China IRA/IHS pair

11.1.1 Coding risk assessments

The aim of coding risk assessments in an IRA is to generate an overall level of risk for each pest associated with the import. To do so, we need to know how risk is assessed. The first step is to identify what risk factors are considered in each assessment. When assessing the risk posed by each pest, the risk factors considered are (Tyson et al., 2009, p. 7):

- Likelihood of entry
- Likelihood of establishment
- Likelihood of exposure
- Likely impacts on the economy, environment, and human health in NZ

The next step is to identify the range of possible likelihoods for each factor. These are:

- Negligible
- Uncertain
- Low
- Moderate
- High

With the factors and their ranges identified we can build a complete picture of how risk is assessed for each pest. This picture can be systematically assembled using the coding scheme in **Error! R eference source not found.** below.

Table A3 - 1: Codes and categories for analysing risk assessments. The categories are the risk factors that we identified, and under each category are the range of possible likelihoods, which are our codes.

Category	Codes	Score
Entry, Exposure, Establishment, Economic Impact, Environmental Impact, Human Health Impact	Negligible	0
	Uncertain	0.5
	Low	1
	Moderate	2
	High	3

The entirety of the IRA is then coded using the scheme above. Each pest considered in the IRA receives a code and corresponding score for each of the six categories. The overall scores were calculated by taking the product of the likelihoods and the sum of the impact assessments (e.g. entry x exposure x establishment x (economic impact + environmental impact + human health impact)). This level can be anywhere from 0-243, with 243 representing a high level of risk for each category and 0 representing a negligible risk level for each category. See Appendix 1 for the full list of pests and their overall risk level.

11.1.2 Coding decisions

When coding the decisions laid out in the IHS for "Pears from China" there were two aims: first, to record which measures are required for each pest and second, to rank the measures by their stringency.

Required measures in the IHS are *actions* that must be undertaken should pests be detected on the consignment. The IHS itself specifies which action must be taken for each pest that is considered, and so coding was simply a matter of recording what action the IHS assigned to each pest.

Table A3 - 2: Actions taken on interception of pest/contaminant

Action Label	**Action Taken on Interception** (MAF Biosecurity NZ, 2010, p. 17)	**Stringency**
NA	No actions as pest is not regulated	Negligible
0	No action due to low risk pathway	Negligible
1	Removal of trash – pests are associated with other plant parts (e.g., leaves, stems, flowers) and/or soil	Low
2	Treat, resort, reship or destroy	Moderate
2a	Treat, reship or destroy. Suspend pathway	High
3	Reship or destroy. Suspend pathway	Very high

Next each measure was ranked for how stringent it is or how capable it is of managing risk the results of which are also displayed in Table A3 - 2. It's not immediately obvious why, for instance, action 2a is more stringent than 2 so some further explanation is required.

- NA/0 – No action required means that these measures are not capable of managing any risk.
- 1 – Removing trash can only manage the risk of pests that do not live in the fruit itself. This measure could not manage the risk of any pest that lives in the fruit, some of which are high risk.
- 2 – Treating or fumigating a consignment might not get rid of all the pests. Resorting involves visually inspecting the pears and discarded infected ones; however, some pests don't have obvious symptoms and so they would be missed. Depending on the level of infestation, the option is available to reship the consignment or destroy it. While doing so would effectively manage the risk from all pests, the fact that these options are presented alongside others that would not, means that this measure is not capable of managing the risk posed by all pests.
- 2a – Like 2, except the option to resort is no longer available. Instead, there is the requirement to suspend the pathway. This means that no more pears from that source can be imported. This measure is capable of managing the highest risk pests, but only if the consignment is reshipped or destroyed. If it is merely treated, some pests may still enter NZ.
- 3 – Like 2a, except treatment is no longer an option. This measure is extremely stringent as no pests could enter NZ on the consignment, and pears from that source are temporarily banned preventing any future risk.

11.1.3 Analysis

Results that may be indicative of misalignment are highlighted below. To see why they are misaligned, it helps to view them in the context of other pests that are considered.

Table A3 - 3: Cases of misalignment in Pears from China IRA/IHS pair

Pest	Risk Score	Action on interception	Alignment
Tetranychus truncatus – cassava mite	6	1 &/or 2	Aligned
Chrysomphalus dictyospermi – Spanish red scale	8	3	Misaligned
Parlatoria oleae – olive scale	8	1 &/or 2	Aligned
Tetranychus kanzawai – kanzawa spider mite	8	1 &/or 2	Aligned
Harmonia axyridis – harlequin ladybird	12	1 &/or 2	Aligned
Lepidosaphes malicola – Armenian comma scale	12	1 &/or 2	Aligned
Pseudococcus comstocki – comstock mealybug	12	1 &/or 2	Aligned
Pseudococcus maritimus – ocean mealybug	12	1 &/or 2	Aligned
Bactrocera dorsalis – Oriental fruit fly	16	3	Misaligned
Conogethes punctiferalis – yellow peach moth	16	1 & 2a	Aligned
Cydia inopinata – Manchurian fruit moth	16	1 & 2a	Aligned
Pandemis spp. – fruit tree tortrix	16	1 & 2a	Aligned
Adoxophyes orana – summer fruit tortrix moth	24	1 & 2a	Aligned
Amphitetranychus viennensis – Hawthorn spider mite	24	1 &/or 2	Misaligned
Spilonota spp. – Tortricid moths	24	1 &/or 2	Misaligned
Monilinia fructigena – European brown rot	36	3	Aligned

Explanations for the misaligned cases are detailed below:

1. Chrysomphalus dictyospermi – This insect is considered a “high-risk” pest in the IHS and receives the strictest possible measure. Yet the IRA reaches a slightly different conclusion only considering the risk to be slightly higher the average of 6.8. 6 pests have and even higher level of risk yet the actions on interception are less severe.
2. Bactrocera dorsalis – likewise with the pest above, this insect, according to the IHS, is so high risk that all consignments must either come from an established pest free area of treated via cold disinfestation, regardless of whether the pest is intercepted. Yet the IRA concludes that while risky, there are pests that present a greater level of risk to NZ.
3. Amphitetranychus viennensis – This mite is one of the riskiest pests according to the IRA yet the measure required on interception would be the same as if it were the lowest risk pest.
4. Spilonota spp. – Similarly, to 3. these insects are considered high risk in the IRA but receive the same measure as the lowest risk pest.

A common factor in these misaligned cases is that ¾ are insects while the other is an insect like mite. However, this is to be expected as the majority of pests associated with the commodity are insects (49/61). While we are not aware of other common factors in these 4 cases, it is possible something in their nature makes it more difficult to assess the risk they pose.

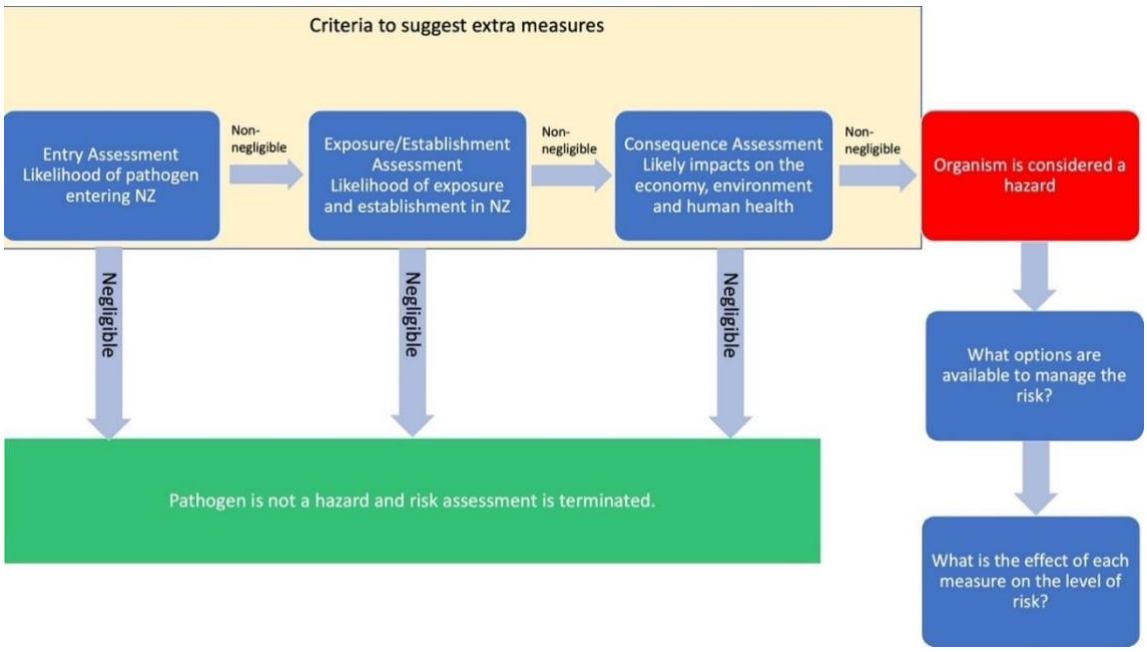
Pears from China overall level of misalignment: 4/33.

11.2 Coding the Malus Nursery Stock IRA/IHS pair

11.2.1 Coding Risk Assessments

To code the *Malus* IRA, for each pest considered, we recorded what pests were considered hazards and for those that were, what measures are suggested to mitigate the risk they pose. To suggest a measure, the IRA first determines if the pest warrants one by determining if the pest is a "hazard". If the pest meets the criteria to be considered a hazard, then the IRA suggests a measure that is appropriate for that type of pest. The risk assessment process is shown in the diagram below:

Figure A3 - 1: Risk assessment process in Malus Nusery Stock IRA. Reconstructed from (Ormsby & Zhu, 2012, p. 5)



To code the IRA, we recorded both the pests that did not meet the criteria to be considered a hazard and those that did. If the organism was considered a hazard, then we also recorded what measure the IRA suggests to manage the risk it poses.

11.2.2 Coding Decisions

Our sample did not include a specific IHS for *Malus*. A specific IHS may exist that was produced shortly after the IRA, however we did not have access to it. Instead, we coded the relevant subsection for *Malus* in the "Importation of Nursery Stock IHS" that covers the importation of a wide range species and was finalised in July 2020.

Measures in the *Malus* subsection are testing requirements that are performed when the consignment is in PEQ. The available tests are:

- Growing season inspection
- Herbaceous/woody indexing
- PCR/ELISA testing
- Herbaceous/woody indexing + PCR/ELISA testing

The aim of coding the relevant subsection of the IHS was to record which of the above measures are required for which pests.

11.2.3 Analysis

To calculate alignment, we compared the results of the risk assessment coding with the results of the decision coding. A decision is aligned with the risk assessment if:

7. No measures are imposed on pests not considered a "hazard"
8. The measure imposed in the IHS matches what is recommended in the IRA

Out of the 40 pests assessed for risk, the following are cases of misalignment.

Table A3 - 4: Cases of misalignment in the Malus Nursery Stock IRA/IHS pair

Pest	Hazard?	IRA Suggested Measures	IHS Imposed Measures	Alignment
Apple latent spherical virus	Yes	Indexing	None	Misaligned
Tulare apple mosaic virus	Yes	Indexing	None	Misaligned
Carnation ringspot virus	Yes	Indexing	None	Misaligned
Tomato bushy stunt virus	Yes	Indexing + Testing	Indexing	Misaligned
Clover yellow mosaic virus	Yes	Indexing + Testing	None	Misaligned
Candidatus Phytoplasma mali	Yes	Testing	Indexing + Testing	Misaligned
Apple bunchy top	No	None	Inspection	Misaligned

Malus Nursery Stock overall level of misalignment: 7/40

11.3 Coding the Rambutan from Vietnam IRA/IHS Pair

11.3.1 Coding Risk Assessments

The aim of coding risk assessments in this IRA is to generate an overall level of risk for each pest associated with the import. In order to do so, we need to know how risk is assessed. The first step is to identify what risk factors are considered in each assessment. The risk factors considered are the very similar to *Pears from China* IRA:

- Likelihood of entry
- Likelihood of establishment
- Likelihood of exposure
- Likely impacts on the economy

The next step is to identify the range of possible likelihoods for each factor. These are:

- Negligible
- Negligible - Low
- Low
- Low - Moderate
- Moderate
- High

With the factors and their ranges identified we can build a complete picture of how risk is assessed for each pest. This picture can be systematically assembled using the coding scheme in **Error! R eference source not found.** below.

Table A3 - 5: Codes and categories for analysing risk assessments. The categories are the risk factors that we identified, and under each category are the range of possible likelihoods, which are our codes.

Category	Codes	Scores
Entry, Exposure, Establishment, Economic Impact, Environmental impact, Socio-cultural impact, Human health impact	Negligible	0
	Negligible – Low	0.5
	Low	1
	Low – Moderate	1.5
	Moderate	2
	High	3

The entirety of the IRA is then coded using the scheme above. Each pest considered in the IRA receives a code and corresponding score for each of the six categories. The overall scores were calculated by taking the product of the likelihoods and the sum of the impact assessments (e.g. entry x exposure x establishment x (economic impact + environmental impact +socio-cultural impact + human health impact)). This level can be anywhere from 0-324, with 324 representing a high level of risk for each category and 0 representing a negligible risk level for each category.

11.3.2 Coding Decisions

When coding the decisions laid out in the IHS for *Rambutan from Vietnam,* there were two aims: first, to record which measures are required for each pest and second, to rank the measures by their stringency.

Required measures in the IHS are steps that must be taken prior to exporting rambutan from Vietnam. They are separated into basic, targeted and MPI-specified measures that are designed to mitigate the risk posed by particular pests.

Table A3 - 6: Measures and their stringency in Rambutan from Vietnam IHS

Measure	Stringency
Basic	Low
Targeted	Moderate
MPI - Specified	High

The variation in the stringency of measures is explained below:

Basic – This measure essentially requires that the fresh rambutan be sourced from a production site that uses cultivation methods for commercial export quality product. This measure is sufficient to manage the risk of the majority of pests by reducing their prevalence in a consignment to very low levels thus limiting their potential to establish and spread if they enter NZ (MAF Biosecurity NZ, 2016, p. 11). This is a low stringency measure because even at very low levels of infestation, certain pests still pose a significant risk which would not be mitigated by the basic measure.

Targeted – A targeted measure is one that is able to manage the risk posed by pests that would not be sufficiently managed with the basic measure. it requires at least one of the following:

- Country freedom from the pest
- Area of production is free from the pest
- Place of production is free from the pest
- In field pest-control activities effective against the pest
- End point treatment effective at managing the pest.

Stringency of a targeted measure is higher than basic because "they provide MPI with the assurance that pest populations on the exported product are reduced to a level that will not enable the pest to establish a population in NZ" (MAF Biosecurity NZ, 2016, p. 12)

MPI-Specified – This measure is reserved for pests that are likely to enter, establish and cause significant impact despite the imposition of targeted or basic measures. The measure requires at least one of the following

- Country freedom from the pest
- Area of production is free from pest
- Consignment is irradiated prior to export.

As we can see, if the country is free from the pest, or the place that rambutan is produced is free from the pest there is little chance it will be present on the consignment. Furthermore, if neither of those criteria can be met, the option is available to irradiate the whole consignment which would destroy the pest if present on a consignment.

11.3.3 Analysis

To code for alignment, we compared the overall level of risk for each pest with the associated measure. Upon comparison we found no instances of potential misalignment.

When the overall level of risk is 0, no measures are required. Basic measures apply to all pests with a level of risk below 37.5, while basic + targeted measures are required for the two pests that score 18. Finally, the riskiest pest receives the most stringent measure.

Rambutan from Vietnam overall level of misalignment: 0/34

11.4 Coding the Prunus Plants for Planting IRA/IHS Pair

11.4.1 Coding Risk Assessments

To code the *Prunus* IRA, for each pest considered, we recorded if it required "extra measures". The minimum requirement for all prunus imports is to spend two growing seasons PEQ where the plants are visually inspected throughout and so "extra measures" are additional tests that must be conducted on the consignment during this period. The risk assessment process that was followed to determine if a pest warranted extra measures is detailed in the figure below.

Figure A3 - 2: Risk analysis process for Prunus Plants for Planting IRA.

As we can see, risk assessment proceeds by asking a series of questions with binary answers. If at any point in the series, the pest answers “No” (or “Yes” in the case of the third question), then the risk assessment is terminated, and it concludes that the pest does not meet the criteria to warrant extra measures. To meet the criteria for requiring extra measures, the analysis must find that the pest is unlikely to be detected by visual inspection alone in PEQ, and it has the potential to cause unwanted impacts in NZ.

To code the IRA, we recorded each pest that met the criteria for extra measures. However, meeting the criteria for extra measures was not only a binary outcome. The IRA found that some pests *may* meet the criteria for extra measures given the uncertainty present in the analysis. The results subsection describes how these cases were handled.

11.4.2 Coding Decisions

To code the decisions in the IHS we needed to determine what measures are prescribed for each pest considered. Measures in this IHS are tests that are carried out when the consignment is in PEQ. The IHS itself provides a list of “Regulated Pests” and their “Mandatory Testing Requirements” (MAF Biosecurity NZ, 2020, pp. 17–18). The

Coding the decisions in the IHS was a matter of transcribing this list into our database.

11.4.3 Analysis

For various reasons not directly relevant to the study, not all pests assessed in the IRA are mentioned in the IHS, and not all pests in the IHS are assessed in the IRA. To code for alignment, we needed to analyse the pests which are assessed in the IRA and receive a measure in the IHS. In total there were 31 pests that were present in both documents.

The cases of clear misalignment are presented in the table below:

Table A3 - 7: Clear cases of misalignment in Prunus for Planting IRA/IHS pair

Pest	Meet criteria for extra measures?	Measures imposed in IHS	Alignment
Phytophthora tropicalis	No	Visual Inspection in PEQ, PCR or Plating onto medium	Misaligned
Phytophthora parsiana	No	Visual Inspection in PEQ, PCR or Plating onto medium	Misaligned

As we can see, the IRA assessed both pests as not requiring extra measures, instead finding that visual inspection alone is sufficient to manage the risk. Nonetheless, the IHS imposed an extra testing measure for each.

The cases of *potential* misalignment are presented in the table below:

Table A3 - 8: Potential cases of misalignment in Prunus for Planting IRA/IHS Pair

Pest	Meet criteria for extra measures?	Measures imposed in IHS	Alignment
Prunus necrotic ringspot virus (almond calico and cherry rugose mosaic strains)	Maybe	Visual Inspection in PEQ, PCR	Misaligned
Nectarine stem pitting-associated virus	Maybe	Visual Inspection in PEQ	Aligned
Phytophthora palmivora	Maybe	Visual Inspection in PEQ, PCR or Plating onto medium	Misaligned
Polystigma rubrum	Maybe	Visual Inspection in PEQ	Aligned

These are potential cases of misalignment because it's uncertain if the pests actually meet the criteria for requiring extra measures. Because of this uncertainty we cannot tell from the IRA alone if the analysis aligns with the decision to impose extra measures or not.

The *Prunus* RMP can offer some insight as the document "gives a rationale for the risk management decisions based on the assessments with significant uncertainty" (Berry et al., 2019, p. 13) like the assessments above.

- Prunus necrotic ringspot virus (almond calico and cherry rugose mosaic strains) – The RMP states that visual inspection alone is all that can be done to manage the risk posed by the strains of this virus. Almond calico and cherry rugose strains are particularly risky, but less severe strains exist as well. Crucially, "PCR testing cannot distinguish between…severe and mild strains" (MAF Biosecurity NZ, 2019, p. 38). Since PCR testing would be ineffective, it is surprising that the IHS decides that it should be mandatory. We might conclude therefore, that this is in fact a case of misalignment.
- Nectarine stem pitting associated virus – The RMP suggests that the impacts of the of the virus are uncertain and "additional measures in post entry quarantine are not justified". Given that the IHS does not mandate any additional measures, we can conclude this case is aligned.
- Phytophthora palmivora – For this pest, the RMP states that since visible symptoms of the pest are likely to develop in PEQ, there is no need for additional testing. Since the IHS requires additional testing, we can conclude that this is a case of misalignment.

- Polystigma rubrum – The RMP finds that this pest "would be appropriately managed in PEQ without any risk management measures other than growing season inspection". Since this is all that is imposed by the IHS, we could say this case in aligned.

Prunus Plants for Planting overall level of misalignment: 4/31

11.5 Full Results of Alignment Coding

Results of alignment coding can be found on the OSF Repository here: https://osf.io/3vryg/