CHARLES UNIVERSITY

FACULTY OF SOCIAL SCIENCES

Institute of Political Studies (IPS)

Department of Security Studies

Master's Thesis

2024 **Alexander Neuhalfen**

CHARLES UNIVERSITY
FACULTY OF SOCIAL SCIENCES

Institute of Political Studies (IPS)
Department of Security Studies

MASTER'S THESIS

Dealing with Uncertainty in Security Analysis Products:
How different Approaches to communicating Uncertainty affect
the Perception of Intelligence Reports and Policy Memos

Author of the thesis:	Alexander Neuhalfen
Study programme:	International Security Studies
Supervisor:	RNDr. Jan Kofroň, Ph.D.
Year of the defence:	2024

Declaration

1. I hereby declare that I have compiled this thesis using the listed literature and resources only.
2. I hereby declare that my thesis has not been used to gain any other academic title.
3. I fully agree to my work being used for study and scientific purposes.

In Prague on 31 July 2024 Alexander Neuhalfen

Bibliographical Note

Neuhalfen, A. (2024). *Dealing with Uncertainty in Security Analysis Products: How different Approaches to communicating Uncertainty affect the Perception of Intelligence Reports and Policy Memos*. Master's thesis (MA). Charles University, Faculty of Social Sciences, Institute of Political Studies, Department of Security Studies. Supervisor RNDr. Jan Kofroň, Ph.D.

Length of the thesis:
157,683 characters with spaces, excluding abstract and reference list

Abstract

Uncertainty is an inherent feature of international relations – particularly, when it comes to matters of national security. Decision-makers are forced to make highly consequential decisions in an environment of incomplete information. It is the job of analysts, e.g. in the intelligence community, to provide them with relevant analyses to support decision-making. But also analysts can't eliminate uncertainty; often, estimative judgements have to be made.

This thesis compares verbal and numerical formats of uncertainty communication and their association with preferences and perceptions of both producers (analysts) and consumers (decision-makers) of analysis products. It does so by conducting an experiment (*N* = 153) which puts participants in the role of both the producer and the consumer of estimative judgements on matters of international security, eliciting their preferences and perceptions in both settings.

The results show a significant shift between producer and consumer preferences towards the numerical format. Numeric precision seems to be particularly demanded in high uncertainty assessments. The data further suggests that numeric probabilities do not create a (false) perception of expertise. However, estimate producers were inconsistent in translating verbal expressions of likelihood and analytic confidence into numeric probabilities.

Given the reluctance of intelligence communities to adopt more precise, numeric formats, this thesis suggests using numeric probabilities strategically – only where it is most needed, to add clarity to the most ambiguous verbal expressions.

Keywords: Intelligence, Intelligence Reports, Policy Memos, Uncertainty, Foresight, Intelligence Estimates, Communication, Perception, Decision-making

Table of Contents

Introduction

Uncertainty is an inherent feature of international relations (e.g. see Rathbun, 2007), particularly, when it comes to matters of international security. None of the actors have complete information on the true intentions and capabilities of the other actors. This uncertainty is even exacerbated when dealing with strategic questions pertaining to the future. Yet decision-makers are forced to make – often highly consequential – decisions in this environment of uncertainty and incomplete information.

It is the job of analysts, for instance in the intelligence community (IC), to provide decision-makers with relevant and actionable analysis, helping them reach well-founded decisions (e.g. see Kent, 1964; Lowenthal, 2006; Jervis, 2010). But even the most advanced of intelligence agencies are not omniscient. Despite all efforts and resources poured into gathering, processing and analysing information, there almost always remains *some* degree of uncertainty. Either, because information is hard to obtain (e.g. state secrets of an adversary) and thus scarce and incomplete, impossible to obtain (e.g. the thinking and true intentions of an authoritarian leader) or even impossible to definitively know (e.g. future events).[1]

> „[M]any of the things you most wish to know about the other man are the secrets of state he guards most jealously. To the extent his security measures work, to that extent your knowledge must be imperfect and your statements accordingly qualified by designators of your uncertainty."
>
> **Sherman Kent, then-head of CIA's Office of National Estimates (1964: 50)**

Unfortunately, the most relevant and consequential questions in national security are often those that are the most uncertain. As General Michael Hayden, former head of

[1] And that only considers the 'known unknowns' – there always remains the possibility of 'unknown unknowns' affecting the whole situation, which further increases uncertainty (see Rumsfeld, 2002).

the NSA and CIA, once succinctly put it: “If it were a fact, it wouldn't be intelligence” (Woodward, 2004: 219). In those cases, analysts often have to make *estimative judgements* – conclusions of analyses made under considerable uncertainty.

But how should analysts deal with that uncertainty in the intelligence products they provide to decision-makers? How should they qualify the amount and the kind of uncertainty their analyses contain? Should they even go so far as to quantify uncertainty with probability values?

What, at first glance, might seem like a minute detail, a mere cosmetic issue of communication, has in fact been the source for much debate over the last half century in western Intelligence Studies (as the literature review will show). Communication *does* matter a great deal in intelligence. *Communicating* the analysis to the decision-maker is, in fact, the bedrock of the dissemination phase of the intelligence cycle.

In 1961, the US Joint Chiefs of Staff concluded in a report that a planned CIA-backed operation to overthrow Fidel Castro had a “fair chance of ultimate success” (Joint Chiefs of Staff, 1961).

> “Despite the shortcomings pointed out in the assessment, the Joint Chiefs of Staff consider that timely execution of this plan has a ***fair chance of ultimate success*** and, even if it does not achieve immediately the full results desired, could contribute to the eventual overthrow of the Castro regime.”
>
> **Joint Chiefs of Staff: “Military Evaluation of the CIA Paramilitary Plan—Cuba” (1961: 69)**

Although there was serious internal debate about the substance and the wording of this assessment, the report was submitted to President Kennedy in early February of 1961 (Bates/Rosenbloom, 1998: 5). The President and many of his staff understood “a fair chance” to be a rather favourable assessment, an endorsement of the CIA’s plan (Friedman, 2019: 3). The authors of the report, on the other hand, later claimed that this assessment was, in fact, meant as a word of warning. “A fair chance” was supposed to reflect a success-probability of 30% – which means a likelihood of non-success, i.e. failure, of 70% (Wyden, 1979: 89). Not aware of that ‘true’ meaning of the assessment, President Kennedy authorised the Bay of Pigs invasion which soon turned out to be a complete disaster.

This anecdote highlights multiple facets and challenges in communicating uncertainty. First and foremost, it illustrates what will later be called *the specificity problem* – how can we ensure that what the recipient of an assessment (the decision-maker) understands is what the producer of the assessment (the analyst) meant? That question inevitably leads to the debate over words or numbers. A discussion that has been going on from the 1950s to this day.

Secondly, it shows that any assessment carries multiple kinds of uncertainty that are relevant for a decision-maker to know. “A fair chance of success” didn’t say anything about how certain the Joint Chiefs were of their assessment. In fact, the authors of the report were “surprised and shocked” given the CIA’s briefing on the operation in the lead-up to the assessment: They expected a detailed plan “thick with documents and appendices” – instead, they were met by six CIA officials who had “brought not a single piece of paper into the room except for a map of Cuba” (Wyden, 1979: 88). The assessment of the Joint Chiefs of Staff therefore rested on a number of assumptions, preconditions and third-party judgements, paired with very limited information (Bates/Rosenbloom, 1998: 5). The report does acknowledge some of these preconditions and assumptions, but it does not communicate clearly how uncertain the Joint Chiefs were of their assessment. It lacked a clear statement on, what today is called, analytic confidence.

These first two facets of uncertainty communication both pertain to informational efficiency – how can information on uncertainty be conveyed to the decision-maker most precisely and clearly? But the example of the failed Bay of Pigs invasion also highlights a third important aspect: In their quasi-symbiotic relationship,[2] analysts and decision-makers are political actors. Thus, from the analyst's point of view, maximum specificity and clarity are not necessarily the only optimisation criteria to strive for. Political considerations – saving face, maintaining credibility and reputation, diffusing blame, avoiding accountability – are also highly relevant in this context, especially when an assessment turns out to be wrong.

That creates the fundamental conundrum that this thesis investigates: In theory, uncertainty communication should be optimised for informational efficiency – i.e. maximum specificity and clarity. That is also what most of the academic literature and intelligence analysis textbooks recommend. However, reality does not reflect that: Intelligence communities are reluctant to adopt measures and standards that would provide more specificity and clarity (Dahmi et al. 2015; Barnes, 2016; Friedman et al., 2018). The thesis tries to better understand the reasons for this divergence between the theoretically optimal (maximum clarity and precision) and the actual practice (an apparent preference of the IC for vagueness).

Finally, the Bay of Pigs anecdote illustrates a fourth point that is relevant to this research: The challenges of communicating uncertainty are not a problem exclusive to intelligence agencies. It is a challenge that all producers of analyses on matters of national and international security face. As Jeffrey Friedman notes in his 2019 book *War and Chance – Assessing Uncertainty in International Politics*: "similar issues recur across a broad range of foreign policy agencies, in public debates among scholars and pundits, and among decision makers forming policy at the

[2] Intelligence producers depend on decision-makers to trust them, to value their analyses and to therefore allocate resources (i.e. budget) towards the IC. Decision-makers, on the other hand, rely on the IC to provide them with relevant, timely, clear and correct analyses that help them make well-informed decisions. For a more thorough discussion of intelligence-policy relations, and an argument why the dependency might be rather lopsided, see chapter two of Rovner (2017).

highest levels." (Friedman, 2019: 49). Therefore, this research is rooted in Intelligence Studies but broadens the scope to include any analysis products on international security. It does so by conducting an experiment which puts participants in the role of both the producer and the consumer of estimative judgements on matters of international security.

This research seeks to help better understand intelligence as a process at its dissemination phase and develop a more comprehensive explanatory framework for the producer/consumer relationship between analysts and decision-makers.

1. Literature Review

This review of the academic literature on uncertainty communication, particularly in the field of intelligence and national security, follows the central conceptual themes that have been debated in post-1945 Intelligence Studies. As an academic field, Intelligence Studies is quite eclectic, interdisciplinary and fairly young – some might also say ill-defined. It has been heavily influenced by other disciplines such as history, political science and psychology.[3] For that reason, this review will feature a lot of relevant literature across academic fields – most notably from psychological research, particularly on human judgement and decision-making. Although this literature review does begin with ideas from the 1950s and 60s and then progresses to more recent research, it does not follow a chronological order. Instead, this review is concept-driven – seeking to retrace the conceptual evolution of and debates around uncertainty communication in intelligence and national security.

The review will start out with the oldest and most fundamental conundrum: *The specificity problem* (or rather: the problem of unspecificity). Then, the review moves to the central question that arises from that; a question which intelligence agencies, decision-makers and scholars have thought and written about for more than 70 years now: *Words or numbers?* That leads to the recommendations that the academic literature has come up with – part three: *Dissecting uncertainty*. However, intelligence agencies have been reluctant – sometimes even resistant – to adopt these recommendations for reform. The reasons for that will be touched upon in part four: *Preferences and Politics*. Whereas part four mainly explores the research on the producer's side of uncertainty communication, the last part contrasts that by

[3] e.g. Richards Heuer's seminal textbook "Psychology of Intelligence Analysis" (1999).

focussing on the intelligence consumers and their *perception of uncertainty,* depending on different formats of communication.

1.1 The Specificity Problem

The problem of specificity is a problem of the delta between meaning and understanding. Simply put: Is what the consumer of an intelligence product understands, exactly what the producer of the report meant? For instance, does the probability that an analyst had in mind when writing about a "serious possibility" align with what a decision-maker reading the report understands to be a "serious possibility"? Sherman Kent faced precisely that question in 1951.

> „Although it is impossible to determine which course the Kremlin is likely to adopt, we believe that the extent of Satellite military and propaganda preparations indicates that an attack on Yugoslavia in 1951 should be considered ***a serious possibility***."
>
> **US National Intelligence Estimate on Yugoslavia, NIE-29 (CIA, 1951: 2)**

Back then, Kent was the vice chairman of the Board of National Estimates within the US intelligence community. The assessment on a possible Soviet attack on Yugoslavia with this key judgement – that "a Soviet attack on Yugoslavia in 1951 should be considered a serious possibility" – was submitted to the US State Department's Policy Planning Staff. Days after receiving the briefing, the chairman of the Policy Planning Staff asked Kent informally: "By the way, what did you people mean by the expression 'serious possibility'? What kind of odds did you have in mind?" (Kent, 1964: 52). For Kent, the expression meant a likelihood of around 65%. The chairman was surprised – he and his colleagues had interpreted "a serious possibility" to correspond to a considerably lower chance. Afterwards, Kent also asked his colleagues at the Board of National Estimates how they interpreted the term: "It was another jolt to find that each Board member had had somewhat different odds in mind and the low man was thinking of about 20 to 80, the high of 80 to 20. The rest ranged in between." (Kent, 1964: 52). According to Kent's own account, that episode prompted him to rethink how analytical products should communicate probability.

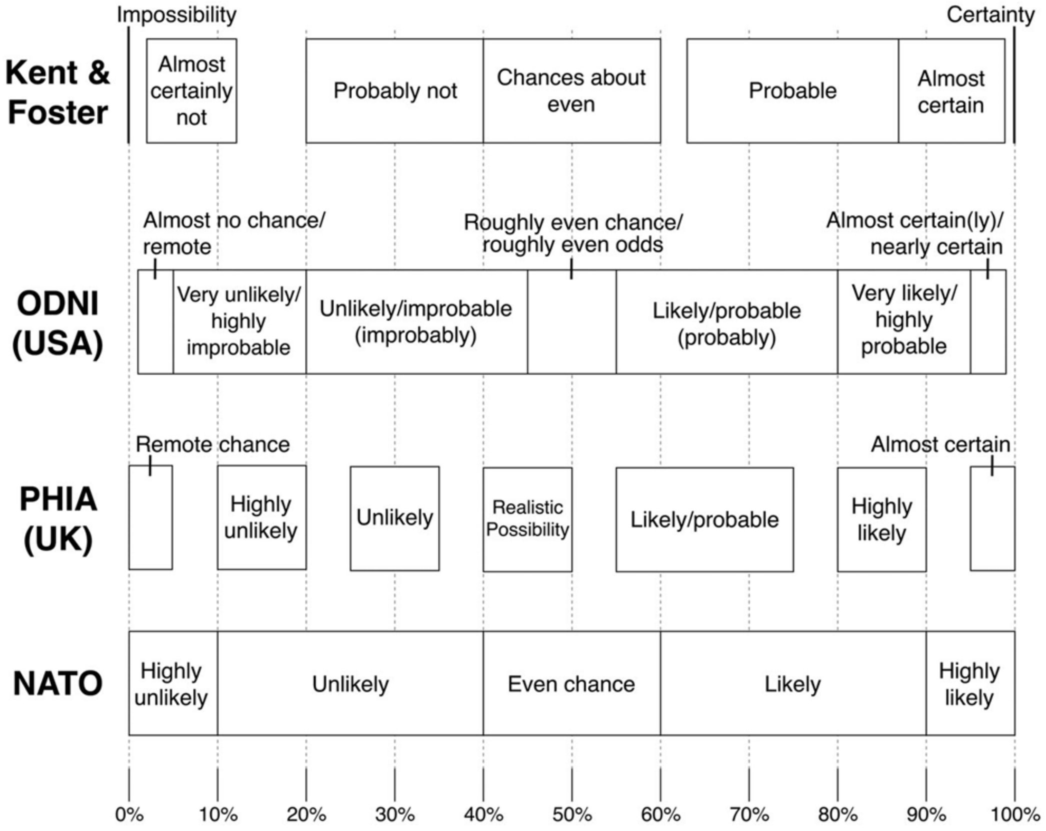


Figure 1.1: *NBLP schemes of different IC's – the original Kent/Foster scheme, the US Director of National Intelligence's (ODNI) current standard, the UK's Professional Head of Intelligence Assessment (PHIA) probability yardstick and the NATO standard (Mandel/Irwin, 2021: 560).*

What he, and his colleague Max Foster, came up with were the *Words of Estimative Probability*; a catalogue of standardised expressions that corresponded to specific numerical ranges of likelihood. Today, such standards are called numerically bound verbal probability schemes, or NBLP schemes for short, and many intelligence communities have developed their own (Fig. 1.1).[4]

[4] In 2019, NATO's *SAS-114 Research Task Group on Assessment and Communication of Uncertainty in Intelligence to Support Decision Making* collected and compared the NBLP schemes of a number of NATO countries (US, UK, Canada, Norway, the Netherlands, and Denmark) as well as from NATO itself (Irwin/Mandel, 2019; NATO STO, 2020).

While intelligence communities have *set* standards for what probabilities certain words correspond to, psychologists have been trying to investigate how verbal probability expressions (VPE) are naturally interpreted. One of the earliest of such studies was conducted by Lichtenstein and Newman in 1967, where participants had to give numerical estimates for “probability-related words and phrases”. In the 1980s, the subject gained more academic attention and since then, a whole host of studies on the numerical interpretation of VPEs has been published (Damrosch/Soeken, 1983; Budescu/Wallsten, 1985; Kong et al., 1986; Brun/Teigen, 1988; Reagan et al., 1989; Hamm, 1991; Sutherland et al., 1991; Tavana et al., 1997; Hobby et al., 2000; Teigen, 2001; Bergenstrom/Sherr, 2003; Honda/Yamagishi, 2006; Cohn et al., 2009; Teixeira/Fialho Silva, 2009; Villejoubert et al., 2009; Juanchich/Sirota, 2013; Shying, 2013; Ostermann et al., 2018). Vogel et al. (2022) did a meta-analysis on these studies and consolidated them into a general overview of the interpretations of VPEs (Fig. 1.2).

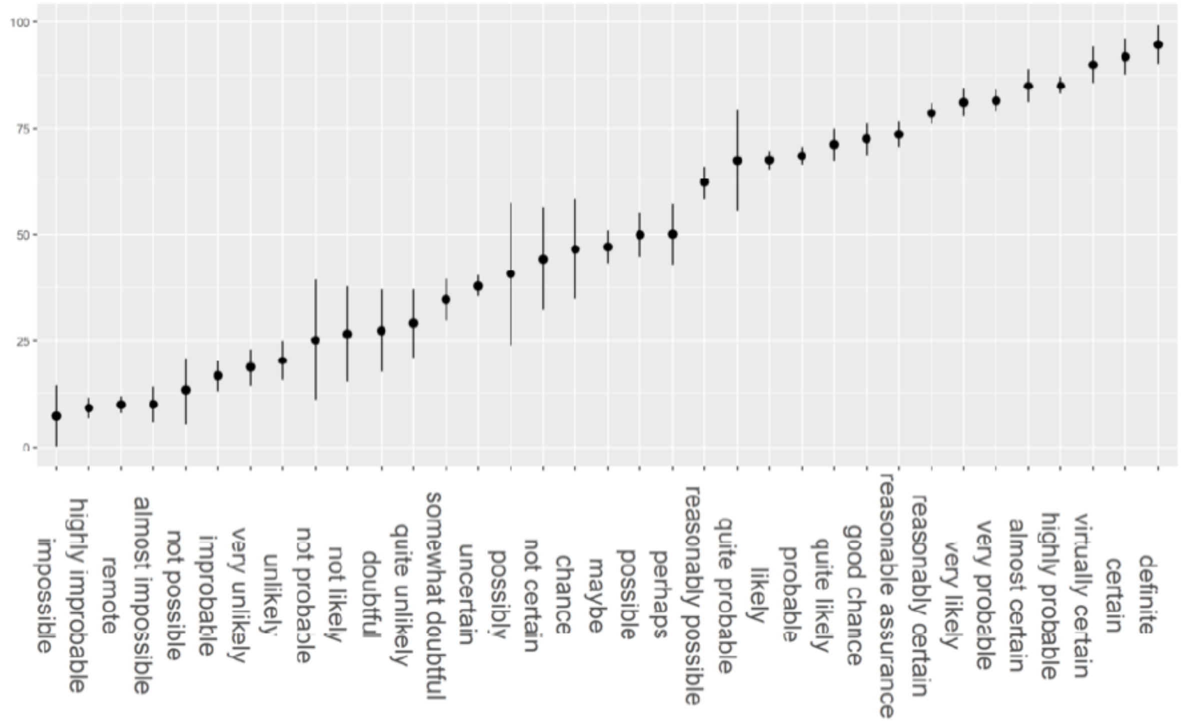


Figure 1.2: *The Interpretation of Verbal Probabilities – averaged mean values of 35 verbal probability expressions; error bars denote the 95% confidence intervals (Vogel et al., 2022).*

These studies showed that VPEs are not interpreted consistently across different individuals (for a more detailed discussion, see for example Dhami/Mandel, 2022) – somewhat supporting the anecdotal evidence Sherman Kent gathered when asking his colleagues about what “a serious possibility” meant.

However, both academic research as well as accounts from intelligence professionals strongly suggest that NBLP schemes do not properly mitigate that problem of inconsistent VPE interpretation. These standards hinge on the precondition that intelligence consumers do not use their own (inconsistent and fuzzy) interpretations of the VPEs but that they adopt the NBLP scheme of the organisation issuing the report. This seems not to be the case: Charts outlining the used NBLP schemes (as e.g. in Fig 1.3) are often ignored or not properly taken into account by intelligence consumers reading and interpreting the analysis (Mandel/Irwin, 2021).

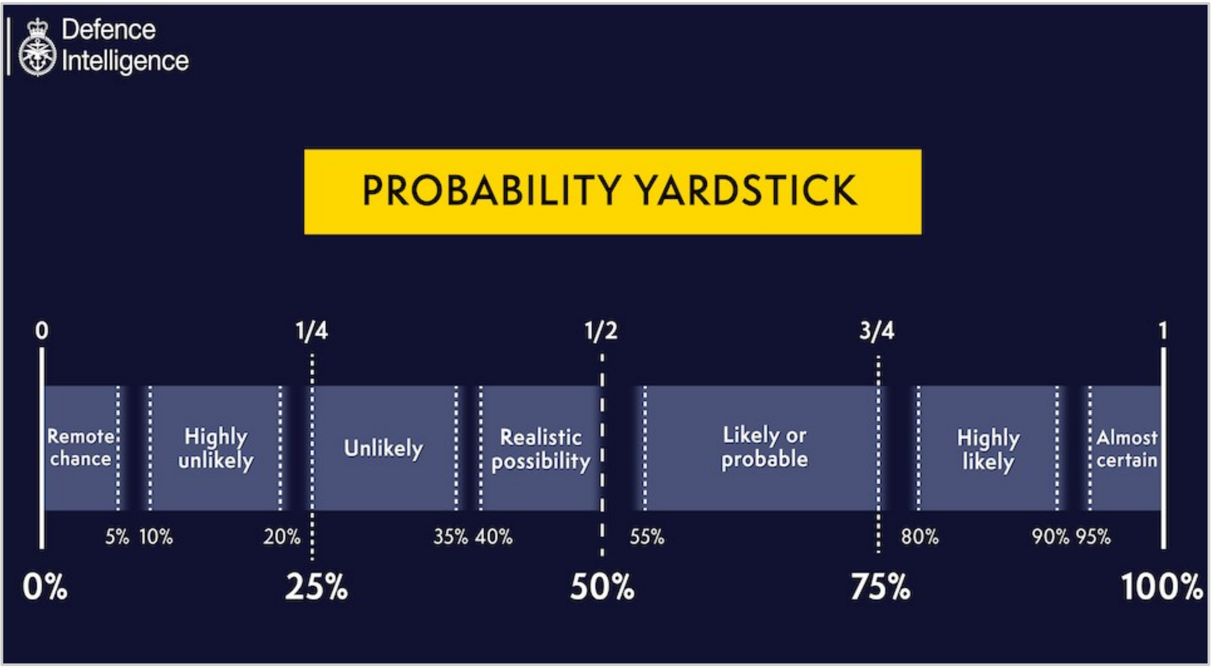


Figure 1.3: *Infographic from the UK’s Defence Intelligence website outlining the PHIA probability yardstick which is used in the public communication of assessments on Russia’s war in Ukraine (Defence Intelligence, 2023).*

As Gregory Treverton, Chair of the US National Intelligence Council from 2014 to 2017, noted: “No policy official ever noticed the chart, much less used it, but it was good discipline for the analysts” (Treverton, 2022: 46).

But even the disciplinary effects of NBLP schemes on the intelligence producers can be questioned. Dhami (2018) found that intelligence analysts had, on average, eight VPEs in their individual lexicon that they used to express probability[5]. And, similar to the natural interpretation of VPEs, these individual lexica were inconsistent across analysts – i.e. different intelligence analysts might very well use different VPEs to describe one and the same probability (Dhami, 2018). Additionally, Ho et al. (2015) found that analysts’ individual VPEs did not necessarily align with their respective agencies’ lexica. That fits Kesselman’s (2008) research of National Intelligence Estimates from the 1950s to the early 2000s which concluded that, despite all efforts of standardisation since Sherman Kent’s first NBLP scheme, there was “a lack of consistency in estimative word usage throughout the years” (Kesselman, 2008: 68).

1.2 Words or Numbers?

Overall, the scholarly literature on communicating uncertainty in intelligence tends to recommend numeric values (e.g. point estimates or, alternatively, probability ranges) rather than VPEs (e.g. Barnes, 2016; Dhami/Mandel, 2021). Or, at least, to complement VPEs with numeric values (Heuer, 1999).

That is due to the many sources of imprecision and miscommunication that VPEs introduce. Imprecisions that are even amplified when multiple intelligence estimates that use VPEs are aggregated into more general judgements – potentially leading to intelligence failure. As the Congressional report on the US intelligence community’s

[5] Based on methods to elicit individuals’ personal VPE lexica that were developed and tested by Budescu and Wallsten (1995), Karelitz and Budescu (2004), Dhami and Wallsten (2005), and Wallsten and Jang (2008).

prewar assessments on Iraq called it: “a ‘layering’ effect whereby assessments were built based on previous judgements without carrying forward the uncertainties of the underlying judgements” (US Congressional Select Committee on Intelligence, 2004).

Numeric values, on the other hand, mitigate the specificity problems of verbal probability expressions. Contrary to VPEs, which typically exist on an ordinal scale (“highly likely” is more likely than “unlikely”, but how much remains unclear), numeric values have ratio scale properties (80% likelihood is twice as likely as 40%). That reduces the potential for misinterpretation, and it allows for aggregating multiple probability judgements (Dhami/Mandel, 2021). Using numeric values also enables agencies to track the performance of their analysts’ estimates, as put forth by Tetlock and Mellers (2011), Mandel and Barnes (2014) and Mandel (2015).

1.3 Dissecting Uncertainty

When talking about uncertainty in intelligence analyses, we are talking about more than just one kind of uncertainty. Friedman and Zeckhauser (2012) criticised the US intelligence community for trying to eliminate or reduce uncertainty in their reports. Instead, they argue that the different uncertainties – emphasis on the plural – should be stated clearly and explicitly in intelligence assessments (Friedman/Zeckhauser, 2015 and 2018). That means, at least, distinguishing between event probability (i.e. the likelihood of whatever is assessed in the judgement) and the analytic confidence of the judgement (i.e. how certain the analyst is of that assessment). In practice, intelligence communities do routinely add an expression of analytic confidence to their assessments – however, the most common format is a coarse verbal ordinal scale of “low”, “moderate” and “high” confidence levels (NIC, 2007; Friedman/Zeckhauser, 2018; Mandel/Irwin, 2021). Moreover, research on nearly 400 declassified US National Intelligence Estimates has found that these two concepts of event probability and analytic confidence are regularly conflated – i.e. analytic confidence is used to express event probability and vice versa (Friedman/Zeckhauser, 2012).

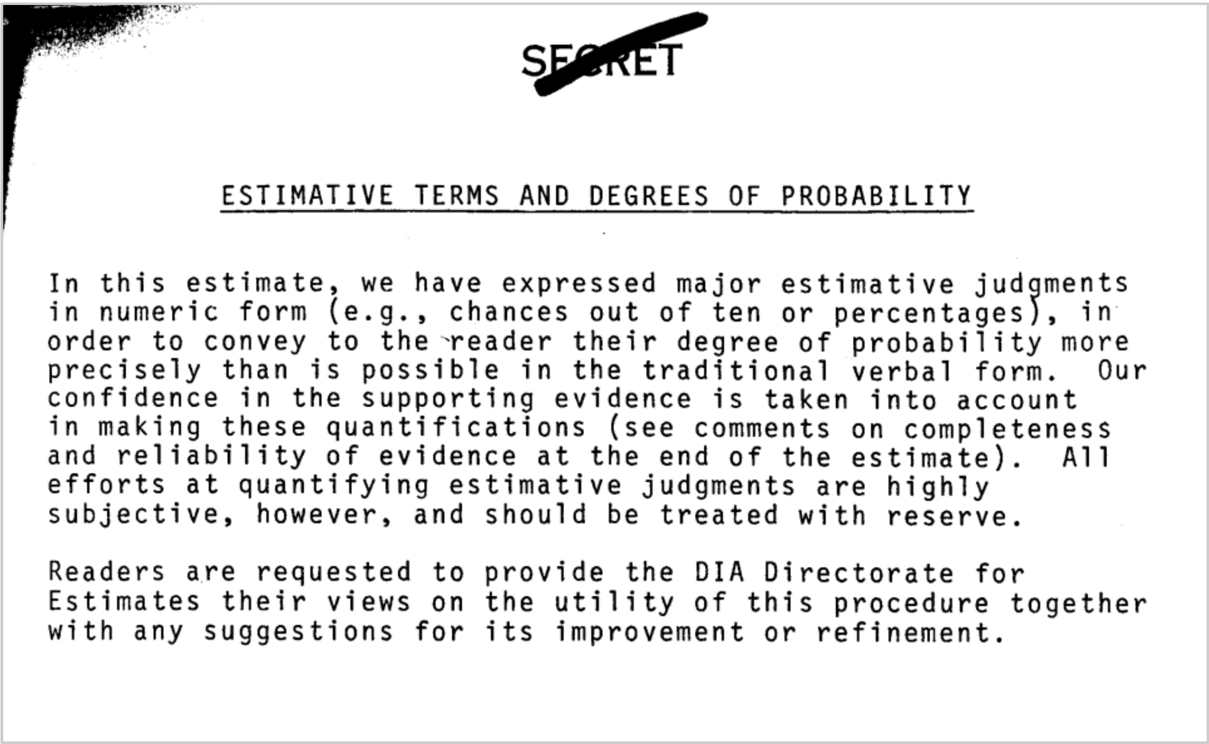
As an alternative to attaching confidence levels, Dhami and Mandel (2021) advocate for using numeric probability ranges when communicating the probability judgement. Either as a point estimate with an added margin of error (e.g. “we assess the likelihood of X to be 70%, plus or minus 10 percentage points”) or as a lower and upper bound of a range (“the likelihood of X lies between 60% and 80%”). Here, the width of the interval implicitly indicates the degree of analytic confidence – the less confident an analyst is about a judgement, the wider the probability range given. Mandel and Irwin (2021) highlight two advantages of that format: First, it expresses the level of analytic confidence as a precise numeric statement and, secondly, it logically links the analytic confidence to the subjective probability judgement.

1.4 Preferences and Politics

Despite most of the academic literature recommending using more precise, consistent, numeric methods of communication, intelligence communities seem to be reluctant to abandon VPEs (Dahmi et al. 2015; Barnes, 2016).

Already in the 1950s and 60s, Sherman Kent noticed a reluctance among analysts to use numeric formats for uncertainty communication. He divided his colleagues into “poets” and “mathematicians” – the *mathematicians* would strive for maximum precision in their words of estimative probability, to the point where they would just use numbers, whereas the *poets* would stick to a more imprecise and inconsistent – but natural – form of communication (Kent, 1964). Given these two factions, Kent couldn’t help but voice his disagreement, and perhaps even frustration, with the *poets*: “They appear to believe the most a writer can achieve when working in a speculative area of human affairs is communication in only the broadest general sense. If he gets the wrong message across or no message at all – well, that is life.” (Kent, 1964: 57).

In later decades, efforts to establish numeric formats have often not seen much success. One illustrative case in point is provided by Marchio (2014): In the 1970s, the US Defense Intelligence Agency (DIA) launched an experiment (Fig. 1.4) wherein they would communicate probabilities in percentages, not VPEs, and always attach statements of analytic confidence to their estimates (A for high, B for medium, and C for low).



~~SECRET~~

ESTIMATIVE TERMS AND DEGREES OF PROBABILITY

In this estimate, we have expressed major estimative judgments in numeric form (e.g., chances out of ten or percentages), in order to convey to the reader their degree of probability more precisely than is possible in the traditional verbal form. Our confidence in the supporting evidence is taken into account in making these quantifications (see comments on completeness and reliability of evidence at the end of the estimate). All efforts at quantifying estimative judgments are highly subjective, however, and should be treated with reserve.

Readers are requested to provide the DIA Directorate for Estimates their views on the utility of this procedure together with any suggestions for its improvement or refinement.

Figure 1.4: *Disclaimer at the beginning of the Defense Intelligence Estimate “Military Significance of Soviet Developed Facilities in Somalia” (DIE SOV 2-76), 20 February 1976 (DIA, 1976).*

Despite the numeric format being well received by consumers of the DIA’s intelligence products, the experiment was discontinued after some years following the turnover of agency leadership (Marchio, 2014). Ironically, today, the DIA is particularly outspoken against the use of numeric formats: “DIA does not condone the use of probability percentages in its products to portray likelihood.” (DIA

Tradecraft Note 01-15, *Expressing Analytic Certainty*, 2015 as quoted in Friedman, 2019: 21).

Mandel and Irwin (2021) have theorised about the reasons for this aversion towards numeric probabilities. They argue that intelligence agencies have very rational incentives to choose vagueness over precision, pointing to its “face-management function”. If an analytic product makes the wrong call, vagueness helps diffuse blame. How can you hold an intelligence agency accountable for a wrong judgement if the analysis never made any precise, concrete claims? Just pointing at what *could* happen makes any falsification hard in the first place (Mandel/Irwin, 2021). The agency “saves face” and therefore the ambiguity and imprecision of VPEs – in the literature mostly considered a liability – ironically becomes an asset. This inclination towards vague language is so strong that Mandel and Barnes (2014), when trying to conduct a quantitative analysis on the accuracy of intelligence forecasts, had to exclude a lot of reports due to vagueness.

In a non-intelligence-specific context, Teigen (2022) calls this preference for vagueness “hedging”, making reference to *Linguistic Politeness Theory*. Speakers deliberately use more vague language to ‘sugar-coat’ (negative) messages (also see Juanchich/Sirota, 2013 and Holtgraves/Perdew, 2016). The IC’s job is not to tell decision-makers what they *want* to hear but what they *ought* to hear (e.g. Gates, 1992).[6] Vagueness might help sugar-coat bad news or make an uncomfortable intelligence assessment more palatable for the decision-maker. To what extent this actually applies in this context remains to be investigated.

Moreover, imprecision creates much room for interpretation. Infamously, the Oracle of Delphi’s secret to seemingly always correctly predicting the future laid in the fact

[6] which has often caused friction between policymakers and the IC – e.g. see Mark Lowenthal’s (1992) much-quoted article *Tribal Tongues: Intelligence Consumer, Intelligence Producers.*

that her predictions were cryptic and therefore vague (Gaub, 2022). Vagueness increases the chance of ‘getting it right’ because an assessment’s meaning can always be bent towards the actual outcome ex post – a phenomenon called “elastic redefinition” (Piercey, 2009).

Some psychologists have also pointed to the *ease of use* for the receivers of probabilistic information as one of the reasons for why communicators of uncertainties seem to gravitate towards VPEs: “It is claimed that most people understand words better than numbers and typically handle uncertainty by means of verbal expressions and associated rules of conversation rather than by numbers” (Budescu et al., 1988: 281).

Overall, the academic literature suggests a certain dynamic: Communicators of uncertainties tend to prefer more vague language, i.e. VPEs. Consumers of probabilistic assessments, on the other hand, rather prefer numeric values. One of the earliest studies showing this was conducted by Wallsten et al. (1993) – almost two thirds of participants preferred to *convey* probabilities verbally, while almost 70% preferred to *receive* probabilities numerically. From Wallsten et al.’s (1993) analysis, parallels can be drawn to Sherman Kent’s conception of *poets* and *mathematicians*: “Generally, respondents who endorsed the use of verbal information said that it is easier to use, as well as more natural and personal. Those preferring numerical information said that it is more precise” (Wallsten et al., 1993: 135).

Irwin and Mandel’s (2023) research puts that into an intelligence-specific context. They investigated the preferences of national security experts and non-experts: Whereas most non-experts favoured the numeric format – similar to Wallsten et al. (1993) found –, the expert group’s preference was evenly split between numbers and

VPEs. However, the majority of both groups – expert and non-expert – perceived the numeric format as more informative.

1.5 Perception of Uncertainty

At that point the debate shifts from a question about preference of the estimate producers to a question about perception of the estimate consumers (or their *anticipated* perception). Two major arguments have emerged in the academic literature to explain estimate producers' preference for more vague formats through the anticipated perception of the estimate consumers.

The first explanation is the *congruence principle*, as proposed by Budescu and Wallsten (1995). This principle states that the preference for communication mode (verbal vagueness or numeric precision) is sensitive to the degree of uncertainty of the underlying assessment. As the Greek philosopher Aristotle (as cited in Friedman et al. 2018) already argued: "the educated person seeks exactness in each area to the extent that the nature of the subject allows". To put that into a modern example: A weather app providing a precise point estimate on the chance of rain tomorrow might be perfectly acceptable; when assessing the likelihood of Iran acquiring nuclear weapons within the next five years, on the other hand, making a precise point estimate might seem misplaced – overly precise, given the uncertainty and unpredictability of that subject matter. The recognition that such precision is unwarranted for such an event may reduce a communicator's perceived credibility (Jenkins et al., 2018). Not being 'overly precise' is often cited as an advantage of verbal formats – Collins et al. (2024) push back at that argument by proposing numeric range estimates, in order to preserve the benefits of numeric precision while also allowing for ambiguity.

Another very common argument is, what Friedman et al. (2017) named, the *illusion of rigour*. Estimative judgements are what decision theorists call 'subjective probabilities' (Friedman/Zeckhauser, 2013). Even if they are made by experts, they remain subjective – based on assumptions and human judgement, not on robust statistical models or even axiomatic deduction. The *illusion of rigour* argument posits that conveying these subjective probabilities in numerical terms creates a false sense of precision and authority (Budescu et al., 1988). One of the standard textbooks for intelligence studies, Mark Lowenthal's *Intelligence: From Secrets to Policy*, also warns: "numerical formulations may be more satisfying than words, but they run the risk of conveying to the policy client a degree of precision that does not exist" (Lowenthal, 2006: 129).

Again, Marchio's (2014) aforementioned account of the DIA's experiment of using numeric probabilities in the 1970s provides a concrete example of that notion. The reason why the DIA experimented with numerical formats was, in part, a dissatisfaction of the Nixon White House with the intelligence community's analytic products. Henry Kissinger reportedly once complained that "analyses and commentaries in the newspapers were superior to anything he read in intelligence publications" (Marchio, 2014: 32). The decision of the DIA to explore quantifying uncertainty can be attributed to the assumption that using numbers instead of words was hoped to boost credibility vis-à-vis intelligence consumers. The subsequent survey among 750 consumers of the DIA's intelligence products seems to confirm that assumption: "A majority favored the use of quantified expressions of probability, believing that they helped to increase their confidence in the information provided and in DIA's judgment and, in particular, helped to give greater credibility to briefings based on the DIA material." (Marchio, 2014: 35). By changing the format of uncertainty communication, the analytic products were perceived as more credible, rigorous and overall more trustworthy.

Friedman et al. (2017) push back on that sentiment, arguing that the numeric format does, in fact, not create a (false) sense of rigour and credibility. In their study, they presented participants (national security professionals and non-experts) with scenarios that demanded a decision under uncertainty – e.g. whether or not to go ahead with a rescue mission in a hostage situation, given that its success is uncertain. They found that participants that received the information about the uncertainties in a numeric format were much more risk-averse (i.e. more hesitant to order the execution of the rescue mission in the hostage scenario) than participants that were given verbal odds. From that, the authors conclude, that numbers have an opposite effect to what the *illusion of rigour* would suggest: Numeric probabilities seemingly heightened the awareness for the degree of uncertainty (Friedman et al., 2017).

In the context of physical and medical risk communication, further research on how different formats of uncertainty communication affect perception has been conducted. Gurmankin et al. (2004) found that numerical formats resulted in higher trust and comfort of subjects in the information they were given, compared to verbal formats. In their 2009 study on communicating cancer risks, Han et al. found that most participants were more worried and perceived the risk to be greater when confronted with a probability range (e.g. 60%-70%) compared to point estimates (e.g. 65%) – hinting at the phenomenon of *ambiguity aversion*.

Longman et al. (2012) compared point estimates versus small and large probability ranges. In line with what Han et al. (2009) concluded, they found that point estimates were better understood and perceived as more credible. On the other hand, more explicit uncertainty communication – by presenting a small or large range of probability – resulted in poorer understanding, increased perception of risk and lowered perceived credibility (Longman et al., 2012). That runs contrary to what the literature from the national security and intelligence context suggests. Dieckmann et al. (2010) presented subjects with fictional intelligence reports in different formats and concluded that subjects were not *ambiguity averse*. Instead, they found the range format to be perceived as “more useful and credible in some cases, particularly in hindsight” (Dieckmann et al., 2010).

1.6 Conclusion

The majority of research on communicating uncertainty in intelligence reports has focussed on the issue of specificity, comparing VPEs to numeric probabilities or investigating differences in individual VPE lexica. From the specificity point of view, the literature is quite clear: Numbers should be preferred over words. Generally, analysts should be as explicit and precise as possible about the different uncertainties by adopting a quantitative approach.

Less attention has been paid to investigating why analysts and intelligence communities might not want to maximise precision. Because, clearly, there is a divergence between the theoretically optimal and the actual practice – which is a reluctance on the part of intelligence producers to provide precise numeric assessments. The congruence principle, as one possible factor causing this divergence, has not been tested in this context. The ‘illusion of rigour’ has been tested in a national security setting, however with a focus only on perceived trust in the estimative judgement itself, not with respect to the perceived expertise of the analyst. Lastly, much of the existing research investigated the preferences and perceptions of intelligence producers and consumers separately. But in reality, most intelligence and national security professionals both produce and consume analyses.

2. Theoretical Framework

This section addresses three core questions in order to embed this research into the existing theoretical landscape. The first question is about the level of analysis – why is this thesis concerned about preferences and perceptions of the individual? The second part comes in from the opposite angle, addresses intelligence's place within the broader framework of neorealist theory: What role does intelligence play in the security dilemma? The third part tries to marry up those first two questions: How does the individual as subject of analysis fit into the theoretical framework of intelligence and its purpose within a state? The key concept to answer that question will be the intelligence cycle. In the fourth and final part, this theoretical reflection will lay out the conceptual framework that underlies this research, based on the theoretical foundations discussed previously.

In its general approach, this research is heavily influenced by the work of Robert Jervis – one of the great theorists in International Relations, who pioneered linking psychological factors, most importantly questions of perception and misperception, to international politics.[7] Jervis is perhaps best known for his work on perception and decision-making in the context of nuclear strategy, deterrence and the security dilemma against the backdrop of the great power-struggle of the Cold War. But Jervis also worked extensively on the study of intelligence.[8] In his 2010 book *Why Intelligence fails: Lessons from the Iranian Revolution and the Iraq War*, Jervis touches on the topics of estimative judgements and uncertainty communication in intelligence (Jervis, 2010: 180ff.).

[7] His seminal work from 1976 is aptly named "Perception and Misperception in International Politics".
[8] e.g. Jervis regularly engaged with the CIA as scientific advisor.

2.1 Do Perceptions matter?

The very first question that Jervis raises in *Perception and Misperception in International Politics* (1976) is also the title of its first chapter: “Do perceptions matter?”. In the face of grand structural theories on the international system[9], on domestic factors[10] and on the dynamics of bureaucracies,[11] why should an individual’s perception play *any* significant role in understanding international politics? Jervis answers that question with a counterfactual: “The three non-decision-making levels [...] say that if we know enough about the setting – international, national, or bureaucratic – we can explain and predict the actor’s behavior.” (Jervis, 1976: 16). But reality shows, Jervis argues, that we can’t explain important events in international politics by just considering these external, structural variables: “it is often impossible to explain crucial decisions and policies without reference to the decision-makers’ beliefs about the world and their images of others. That is to say, these cognitions are part of the proximate cause of the relevant behavior and other levels of analysis cannot immediately tell us what they will be.” (Jervis, 1976: 28). The decision-maker’s perception of the world has to be considered in tandem with the structural factors shaping that world. Just as the international, national and bureaucratic level, decision-making should be considered an equal, fourth level of analysis and thus demands investigating decision-makers’ perceptions (Jervis, 1976).

That’s what this thesis does by looking into the preferences and perceptions of individuals – archetypically named ‘analysts’ and ‘decision-makers’ – and linking that to the structural context they operate in.

[9] Like Neorealism for instance.

[10] e.g. through game theory (see Putnam, 1988).

[11] ranging from Max Weber’s Bureaucracy Theory to rational-choice models from Political Economy.

2.2 The Security Dilemma and Intelligence

Uncertainty lies at the very heart of the security dilemma, one of the foundational concepts in neorealist theory. As Maersheimer (1994: 10) writes: “States can never be certain about the intentions of other states. Specifically, no state can be certain another state will not use its offensive military capability against the first”. To neorealists, that uncertainty drives state behaviour and therefore the dynamics of the international system.[12] In this environment of uncertainty and anarchy, decision-makers are forced to make consequential decisions under incomplete information.

That’s where intelligence comes in. Intelligence exists to help decision-makers navigate this uncertain environment of the international system. As Pashakhanlou (2018: 522) phrases it: “Intelligence is employed to gauge the capabilities and intentions of others in the security dilemma [...] to avoid strategic surprise”. In many countries, key government figures are routinely briefed by their intelligence agencies on relevant issues.

But despite all the advanced means of gathering information and elaborate methodology of analysis, there are still limitations: “intelligence can be an indispensable tool in the security dilemma [...]. Yet, it is not an exact science and the most intelligence can accomplish is to narrow the level of uncertainty and enable more informed decision-making. It cannot eliminate uncertainty, identify each risk or anticipate every contingent scenario.” (Pashakhanlou, 2018: 523).

2.3 The Intelligence Cycle

To combine questions of preference and perception on the individual level with the broader purpose and function of intelligence within a state, this research uses the

[12] for more comprehensive theoretical discussions see Jervis (1978) and Waltz (1979: 186ff.)

perhaps oldest and best-known theoretical framework from Intelligence Studies: The Intelligence Cycle.

Sherman Kent popularised this concept in his seminal 1949 work *Strategic Intelligence for American World Policy*. The intelligence cycle has since become the most common representation of intelligence as a process (Fig. 2.1). It conceptualises the intelligence process as a sequence of stages that form a loop. The process is initiated by decision-makers that formulate tasks and objectives for the intelligence agency – in *Planning and Direction*. That is followed by the operational stages *Collection*, *Processing and Exploitation*, and *Analysis and Production* (which are executed by the intelligence agencies). In the last phase, *Dissemination and Integration*, the Intelligence product gets fed back to the decision-makers, resulting in new or updated directions and orders for the intelligence agencies – and the cycle starts anew.

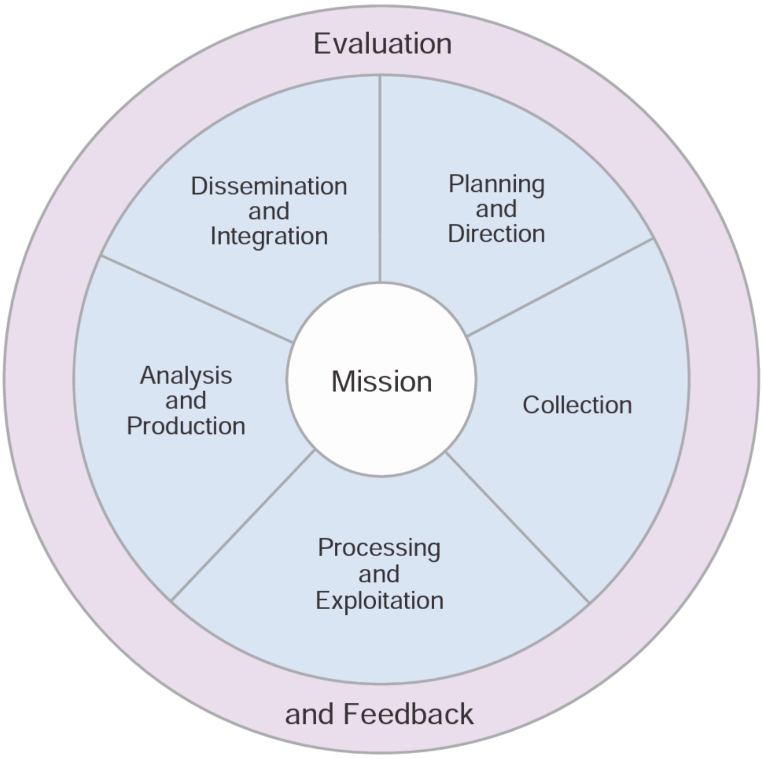


Figure 2.1: *The intelligence cycle as in JP 2-0 (Joint Chiefs of Staff, 2013).*

As early as in 1949, Kent's conceptualisation of intelligence as a cycle – driven by policymakers, with the intelligence agency as a rather passive provider of information – was already criticised.[13] Scrutiny of the intelligence cycle continued throughout the decades. A contemporary critique is its conceptual linearity and sequentiality, rendering it an oversimplified representation that is not capable of accommodating relevant phenomena like covert action, counterintelligence, and oversight (Gill/Phythian, 2013). Gill and Phythian (2013) therefore propose a network rather than a cycle (Fig. 2.2).

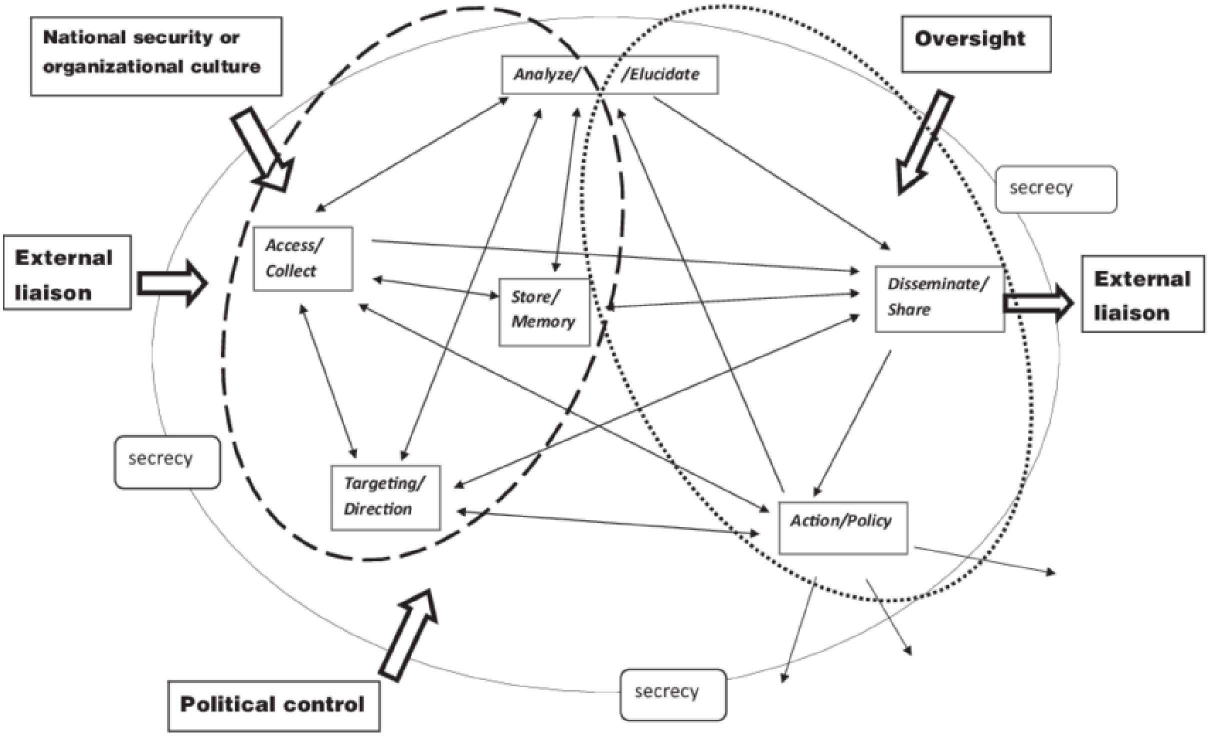


Figure 2.2: *The intelligence web (Gill, 2020: 46).*

Despite all academic debate, it is undisputed that *Dissemination* is an important aspect of intelligence. For instance, it is also still featured in the contemporary network conceptualisation of intelligence as a key element (Fig. 2.2). As a starting point, however, the 'classic' conceptualisation of the intelligence cycle remains

[13] See for example 'the Kent-Kendall Debate' of 1949 (Davis, 1991)

applicable because this research does not investigate the intelligence process as a whole. Instead, it only considers a fraction of the intelligence cycle: The consumer/producer interaction between the analyst and the decision-maker.

2.4 The Producer/Consumer Relationship

> „There is no phase of the intelligence business which is more important than the proper relationship between intelligence itself and the people who use its product.“
>
> **Sherman Kent (1949: 180)**

Analysts and decision-makers interact in a producer/consumer relationship. This relationship is embedded into the broader relationship between decision- and policymaking (e.g. in the government or legislature) and the intelligence community (Fig. 2.3).

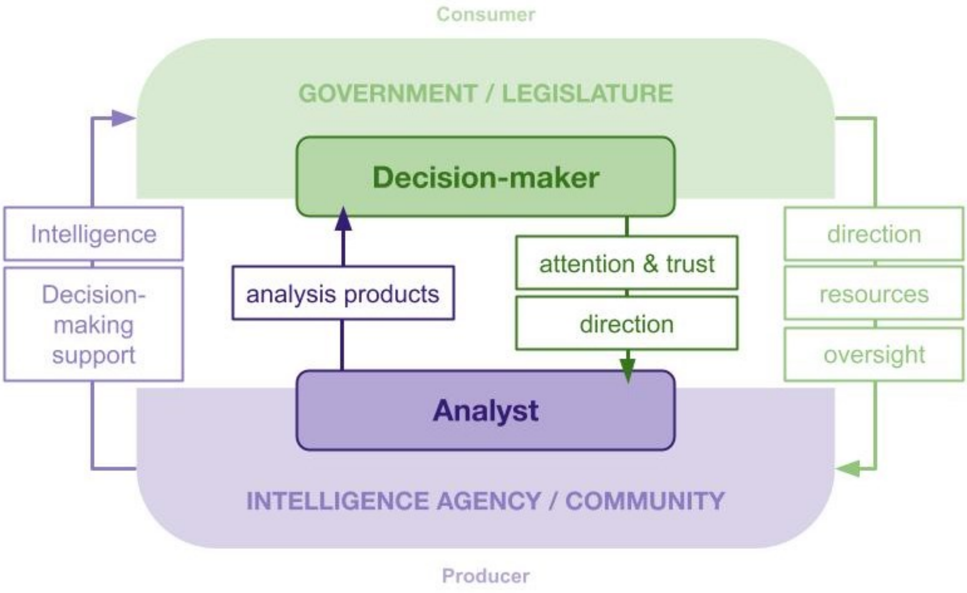


Figure 2.3: *Conceptualisation of the producer/consumer relationship between analysts and decision-makers in an intelligence context (for the purpose of this research)*

In the *dissemination* phase of the intelligence cycle, the analyst provides the decision-maker with analysis products – all within the broader function of the intelligence community to deliver relevant intelligence that, optimally, provides decision-making support (left side of the schematic in Fig. 2.3).[14]

But that relationship is not a one-way-street; it only works if the decision-maker awards attention and trust to the intelligence. Otherwise, intelligence failure can be the consequence (e.g. see Dahl, 2013). If the intelligence is ignored, dismissed or mistrusted by decision-makers, then intelligence agencies also lose their legitimacy, their *raison d'être*. Assuming that the government and/or the legislative – besides oversight functions – hold power over the intelligence community's budget, such a loss in relevance and legitimacy constitute an existential issue. Thus, intelligence agencies depend on being perceived as credible and useful.[15]

Furthermore, and in line with the *direction* phase of the intelligence cycle, the decision-maker influences what kind of intelligence is produced – for example by requesting analysis on certain topics and questions. This can play out on the small scale between decision-maker and analyst, but also on the functional, institutional level (right side of the schematic in Fig. 2.3).

It has to be noted that this conceptualisation depicts a very idealised version of the producer/consumer relationship in intelligence; what it *should* look like – in theory. That reality looks very different is poignantly described in Mark Lowenthal's often cited article *Tribal Tongues: Intelligence Consumers, Intelligence Producers* (1992). For example, when it comes to the 'support function' for decision-making of

[14] Sherman Kent, for example, emphasised this support function of intelligence: "Its job is to see that the doers are generally well-informed; its job is to stand behind them with the book opened at the right page [...]" (Kent, 1949: 182).

[15] As Robert Gates remarked when he was CIA Director: "If we ignore policymaker interests, then our products become irrelevant in the formulation of our government's foreign policies." (Gates, 1992). Similarly, Sherman Kent warned: "Intelligence cannot serve if it does not know the doers' minds; it cannot serve if it has not their confidence" (Kent, 1949: 182).

intelligence (left side of the schematic in Fig. 2.3) – intelligence producers and consumers often have very different views on what ‘good’ decision-making support looks like (Lowenthal, 1992).[16] With regards to the right side of the schematic (Fig. 2.3), there is also much potential for friction[17] and dysfunction. Over the past decades, ‘Politisation’ – “the attempt to manipulate intelligence so that it reflects policy preferences” (Rovner, 2017: 5) – has been the most widely discussed issue and has often been the subject of intelligence reform.

From this conceptualisation (as in Fig 2.3) of the producer/consumer relationship, certain implications for how to handle uncertainty follow: On the one hand, it is the purpose of any intelligence product to shed light into darkness, to reduce uncertainty for the decision-maker. On the other hand, most intelligence products just cannot eliminate uncertainty – even though that might be what some decision-makers demand (Lowenthal, 1992: 15; Rovner, 2017: 4). That is a challenge intelligence producers have to grapple with.

> „While we strive for sharp and focused judgments for a clear assessment of likelihood, we must not dismiss alternatives or exaggerate our certainty under the guise of making the ‘tough calls’. We are analysts, not umpires, and the game does not depend on our providing a single judgment.”
>
> **Robert Gates, then-Director of the CIA, in his address to analysts (Gates, 1992: 9)**

[16] As Lowenthal (1992: 13) writes: “For policymakers. This means a shared and active interest and, if necessary, advocacy. This runs counter, however, to the intelligence community's long-standing position not to advocate any policy. Rather, the intelligence community tends to see itself, correctly or not, as a value-free service agency, although at its upper levels the line begins to blur.”

[17] There are countless examples of top-level decision-makers that thought very little of their respective intelligence agencies in terms of credibility. On the troubled relationship between US presidents and their intelligence agencies see Manjikian (2020); with focus on the Nixon administration Marchio (2014). And as a non-US example: German chancellor Helmut Schmidt reportedly called the BND a “bunch of dilettantes” [“Dilettantenverein”] (Spiegel, 1984).

While this simple conceptualisation already provides a useful framework for this research, it lacks two aspects. First, it is still a bit under-complex: No intelligence product is produced by one ‘rank and file’ analyst and then goes straight up to the head of government. The conceptual framework for this research needs to reflect the different layers on which producer/consumer relationships take place. As Rovner (2017: 19) elaborates: “The interaction between the intelligence and policy communities takes place continually; just as high-level policymakers deal with senior intelligence advisors, policy staffers and intelligence analysts communicate formally and informally at lower levels”.

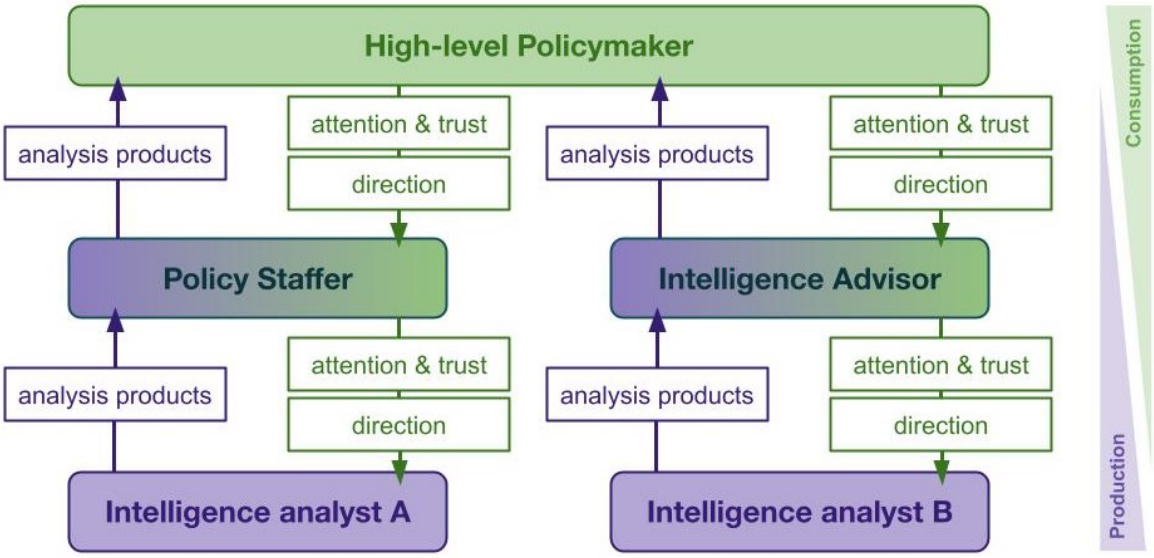


Figure 2.4: *Conceptualisation of the multi-layered producer/consumer relationship in an intelligence context (for the purpose of this research)*

It is rare that a given person *only* consumes or *only* produces analysis products. Such cases would presumably only be found in staff analysts (high production, low share of consumption of intelligence) and top-level decision-makers (little to no production, but frequent consumption of analysis products). In between those two extremes, however, there are many hybrid-positions: Mixed producer/consumer roles, in which a person does receive analyses but still has to report to a superior and also acts as analysis producer (the middle layer in Fig 2.4).[18] This research is sensitive to that complexity by asking about participants' experience as producers and consumers of estimative judgements separately.

If the consideration of multiple layers can be pictured as 'deepening' the initial conceptual framework for this research, then the second amendment would be considered a 'widening'. As argued earlier, challenges of the producer/consumer relationship between analysts and decision-makers are not exclusive to the intelligence context.[19] This research, therefore, considers not only 'classical' intelligence producers and consumers. Instead, the scope is widened to any producers and consumers of estimative judgements on national and international security – within ministries, in the armed forces, between decision-makers and academia, between decision-makers and think tanks, in the private sector, et cetera. All under the assumption that the core qualities of the producer/consumer relationship identified in the intelligence context – the consumer demands relevant analysis to support decision-making; the producer requires the consumer's attention and trust (and direction) – and its challenges to handling uncertainty also apply in these other contexts.

[18] Note that the titles of the positions, such as "policy staffer" or "intelligence advisor", are not exhaustive – i.e. there are many more positions that would be located in that middle layer of mixed producer/consumer roles.

[19] "[S]imilar issues recur across a broad range of foreign policy agencies, in public debates among scholars and pundits, and among decision makers forming policy at the highest levels." (Friedman, 2019: 49)

3. Methodology

The literature strongly suggests that intelligence producers have a preference for more vague verbal formats and/or an aversion towards precise numerical formats to convey estimative judgements. That seems counterintuitive, because it goes against the principle that intelligence analyses ought to communicate uncertainty as clearly and precisely as possible to its consumer. The thesis seeks to better understand the reasons for this divergence between the theoretically optimal (maximum clarity and precision in dealing with uncertainty) and the actual practice (an apparent preference of the IC for vagueness). To that end, the research focuses on three possible reasons that might explain this divergence.

Ontologically, this research adopts a clearly objectivist approach, treating the preferences and perceptions of its subjects as observable and measurable qualities. Through hypothesis testing and quantitative analysis, this research intends to produce insights that are somewhat generalisable beyond the experimental setting. In doing so, it follows a positivist epistemological view.

3.1 Research Hypotheses

The first possible explanation from the literature that will be tested is the ***face-management function*** of verbal vagueness (Mandel/Irwin, 2021). If that were the case within this context, format preference should be highly dependent on the position within the social context and the degree of perceived accountability for the estimate. That is particularly relevant if there is a hierarchical delta between the estimate producer (e.g. an analyst) and the estimate consumer (a superior or a decision-maker).

Hypothesis 1: *When participants feel accountable for an estimate vis-à-vis a superior, they prefer verbal formats whereas they favour numerical formats when personal responsibility and hierarchical pressure is not present.*

The second possible explanatory factor being investigated is the so-called ***congruence principle*** (as proposed by Budescu/Wallsten, 1995). The congruence principle suggests that participants are more inclined towards choosing formats with a higher specificity when the likelihood is either very high or very low and/or when analytic confidence is high. Conversely, if the outcome is deemed to be very uncertain – e.g. when the odds are about 50/50 – participants would be expected to gravitate towards more vague (verbal) formats.

Hypothesis 2: *Estimate producers are more likely to choose numerical formats the more certain the assessment is.*

Somewhat related to the congruence principle is the ***illusion of rigour*** (described by Budescu et al., 1988). If a very precise, numeric format is chosen for something that, naturally, is very uncertain, that might create a false unwarranted sense of precision and certainty. Friedman et al. (2017) dispute that notion, showing that decision-makers even were more risk-averse when presented with numeric odds – the opposite of what would have been expected under the illusion of rigour.[20] But that experimental setting focussed on whether the participants *trusted* the assessments. It did not investigate the participants' perception of the expertise behind the producer of the assessment. But, besides trust, that's an important second component of credibility: expertise (Hovland et al., 1953; Wiener/Mowen, 1986). This research approaches the question whether an illusion of rigour exists in

[20] A more detailed account of the study is provided in *Perception of Uncertainty* in the literature review

the context of intelligence/national security assessments from the angle of this second component: perceived expertise.

Hypothesis 3: *Consumers of estimative judgements associate numerical formats with a higher expertise of the producer behind the assessment.*

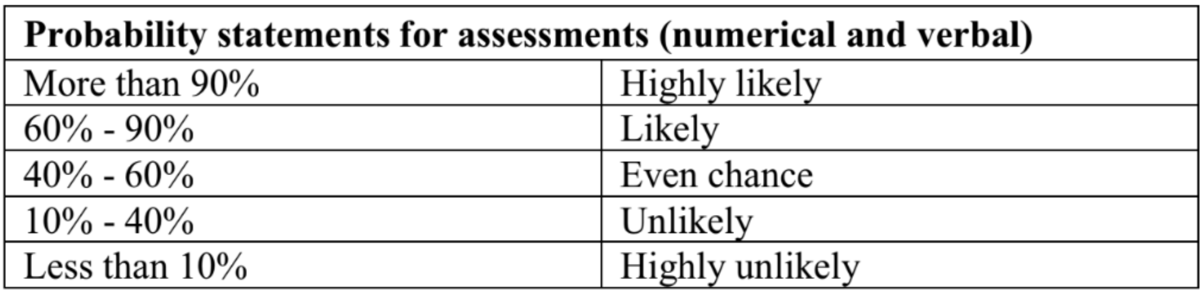
Methodologically, the format to communicate political assessments and the uncertainty they contain plays a central role in this research. For hypothesis one and three, the format to is the dependent variable; for hypothesis two, the format constitutes the independent variable.

3.2 Selection of tested Formats

Since the purpose of this research is to investigate possible differences in perception and preference between verbal and numerical formats, both formats have to be featured. In both cases, however, there exist multiple forms and standards. Even if natural (unstandardised) estimative language is disregarded for this purpose, there still remains a whole host of NBLP schemes as possible verbal formats to be tested. There are several VPE systems to convey likelihood (Irwin/Mandel, 2019) and on top of that there is always the option of adding a verbal qualifier on analytic confidence. For the numeric formats, the possible options are point estimates – in percent or in odds (like “a 7 in 10 chance”) – and range estimates – either as a point estimate with an added margin of error (“70% plus or minus 10 percentage points”) or as the range itself (“likelihood between 60% and 80%”). While range estimates contain information about the analytic confidence of the assessment, ‘pure’ point estimates don’t (Dhami/Mandel, 2021). Between those two groups of verbal and numerical formats, there is also a ‘hybrid’ or ‘mixed’ format of combining VPEs with added numerical point estimates in brackets (e.g. “highly likely [90%]”).[21]

For the numerical formats, this research will utilise range estimates. They convey both an assessment on likelihood and analytic confidence; it is therefore recommended as a format to be used in intelligence by some scholars since it intuitively links these two concepts of uncertainty (likelihood and confidence) that are often conflated (Mandel/Irwin, 2021; Dhami/Mandel, 2021). The estimate will be created as a point estimate with an added margin of error; in the final assessment, though, the estimate will be displayed as a simple range from the lower to the upper limit.

The VPEs used in the survey are based on the NATO NBLP scheme, rather than that of a particular national intelligence community or agency. The NATO standard is outlined in the NATO Allied Joint Doctrine for Intelligence Procedures (NATO AJP-2.1). While the AJP-2.1 itself is not an open source available to the public, its NBLP scheme has been discussed in official NATO documents, such as the report of the *Research Task Group SAS-114* (Fig 3.1).



Probability statements for assessments (numerical and verbal)	
More than 90%	Highly likely
60% - 90%	Likely
40% - 60%	Even chance
10% - 40%	Unlikely
Less than 10%	Highly unlikely

Figure 3.1: *NBLP scheme from NATO AJP 2.1 2016 as cited in Irwin and Mandel (2019: 298)*

For the purpose of this survey, however, one modification to the NATO terminology has to be made: the VPE “even chance” does not work well in the German language

[21] Particularly, this format has been recommended in Richard Heuer’s intelligence textbook “Psychology of Intelligence Analysis” (1999). This format was also featured in research by Jenkins et al. (2018 and 2019).

(which is one of the three languages that the survey will be disseminated in). To express the meaning of “even chance” in German, one would have to revert to an expression like “50/50 chance”, which would introduce a numerical probability expression into a format that is supposed to be a purely verbal one. For this reason, the term “even chance” is omitted from the VPEs used in the survey. That means that participants are presented with the options “highly unlikely”, “unlikely”, “likely” and “highly likely” to express likelihood. Not having a separate expression for even odds and just jumping from the under-50% probability space (“unlikely”) to the above-50% probability space (“likely”) is not uncommon in NBLP schemes. For instance, the VPE standard of the Dutch Defense Intelligence and Security Service (DISS) – a language related to German – does this as well (Fig 3.2).

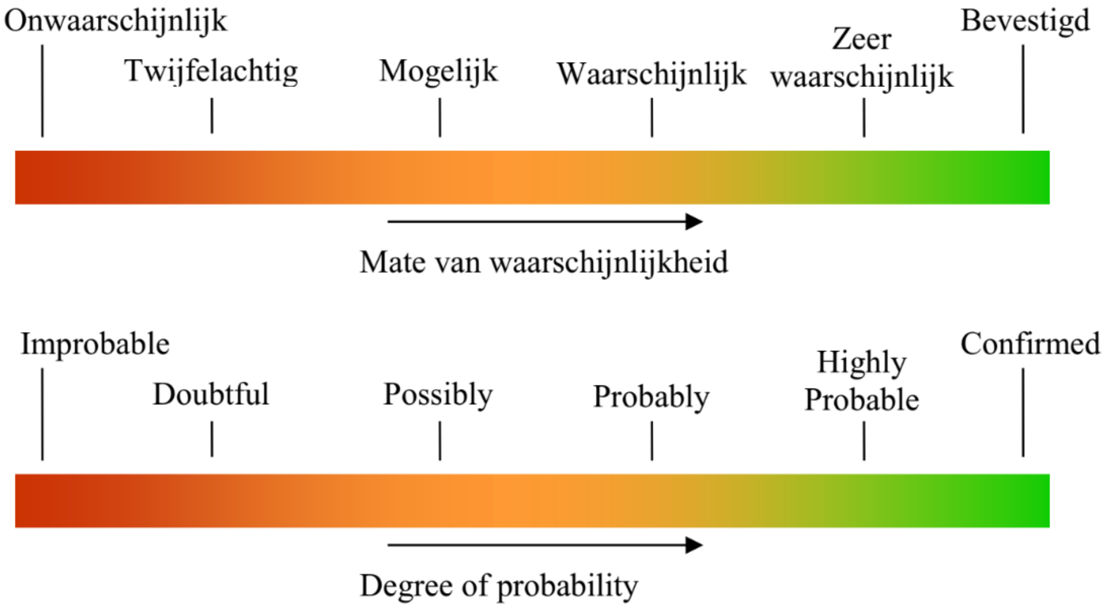


Figure 3.2: *DISS degrees of probability as cited in Irwin and Mandel (2019: 305)*

Also featuring the hybrid format (of a VPE paired with a numerical point estimate) in the survey was considered – as a ‘middle ground’ between the purely verbal and the purely numerical format. In the end, however, this mixed format was not included since it posed the risk of negatively affecting the elicitation of format preference. When asked whether they would prefer numerical or verbal formats, participants might be inclined to ‘play it safe’ and to select the hybrid format that includes both. Instead, a different third format was chosen to be included in the survey: The VPE paired with a verbal statement on analytic confidence.

As mentioned previously, numerical range estimates convey both likelihood and analytic confidence. The survey should offer participants a verbal ‘alternative’ that contains the same amount of information. That means adding an expression of analytic confidence to the VPE on likelihood. To this end, this research opted for the standard of the US intelligence community, which is arguably the most common format: A separate sentence expressing the level of analytic confidence on a three-tiered scale of “low”, “moderate” and “high” (NIC, 2007; ODNI, 2022).

In conclusion, the three formats used in this research are the following:

1. A Verbal Probability Expression (VPE) on likelihood
   (*“It is highly likely that X happens.”*)
2. A VPE on likelihood with an added qualifier of analytic confidence
   (*“It is highly likely that X happens. The analytic confidence in that assessment is low.*”)
3. A numerical range
   (*“The likelihood of X lies between 75% and 95%.”*)

3.3 Survey Design

This research employs quantitative methods to investigate the dynamics between producers and consumers of estimative judgements on national and international security. To empirically collect the necessary data to test the hypotheses outlined in the previous chapter, an online survey is used. The survey consists of four main phases, in each of which participants have to perform different tasks. Between phase two and three, there is a short break where participants are asked to provide some personal information.

Phase I – Self-Evaluation

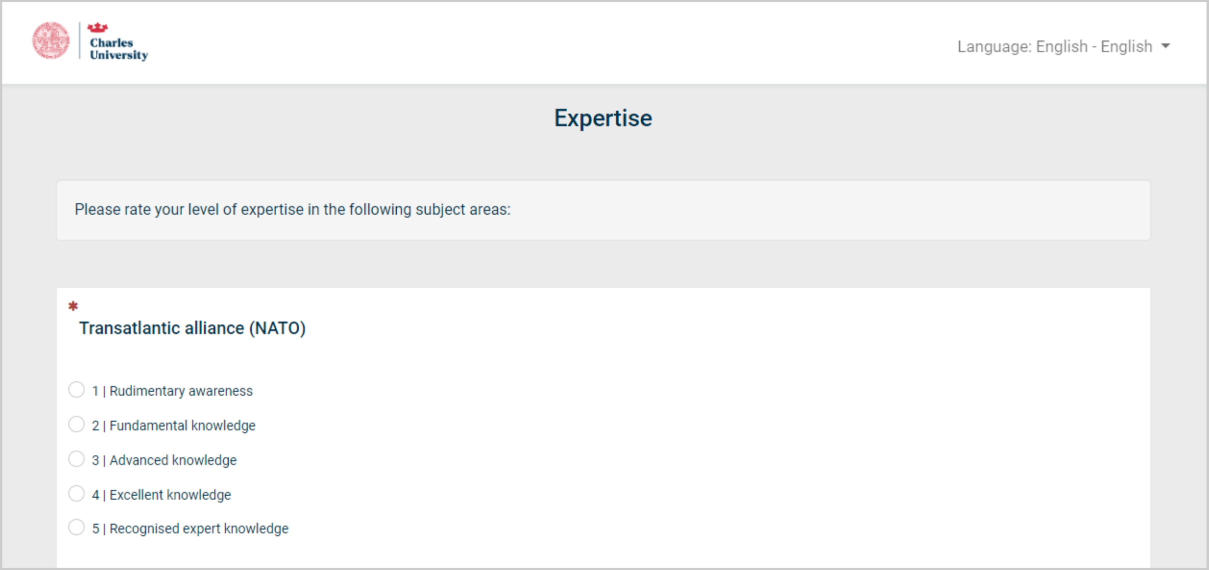
The survey starts out with a self-evaluation of the participants' own expertise and experience. First, participants have to rate their expertise in three subject areas:

1. Transatlantic alliance (NATO)
2. Russia's War in Ukraine
3. East Asian security (China/Taiwan)

For each subject area, participants can select their level of expertise out of five options (Fig. 3.3):

1. Rudimentary awareness
2. Fundamental knowledge
3. Advanced knowledge
4. Excellent knowledge
5. Recognised expert knowledge

Since these options are not further specified or standardised in any way and participants' self-evaluation is highly subjective, this part of the survey does not yield a very solid dataset on the actual expertise of the participants – but that is also not its purpose. Instead, the answers given in the self-evaluation part influence what subject the participants are presented with in phase three of the survey. The results of the self-evaluation feed into a selection algorithm that is optimised to select the subject in which the participant's expertise is 'the most mediocre'.[22]



Expertise

Please rate your level of expertise in the following subject areas:

* Transatlantic alliance (NATO)

○ 1 | Rudimentary awareness
○ 2 | Fundamental knowledge
○ 3 | Advanced knowledge
○ 4 | Excellent knowledge
○ 5 | Recognised expert knowledge

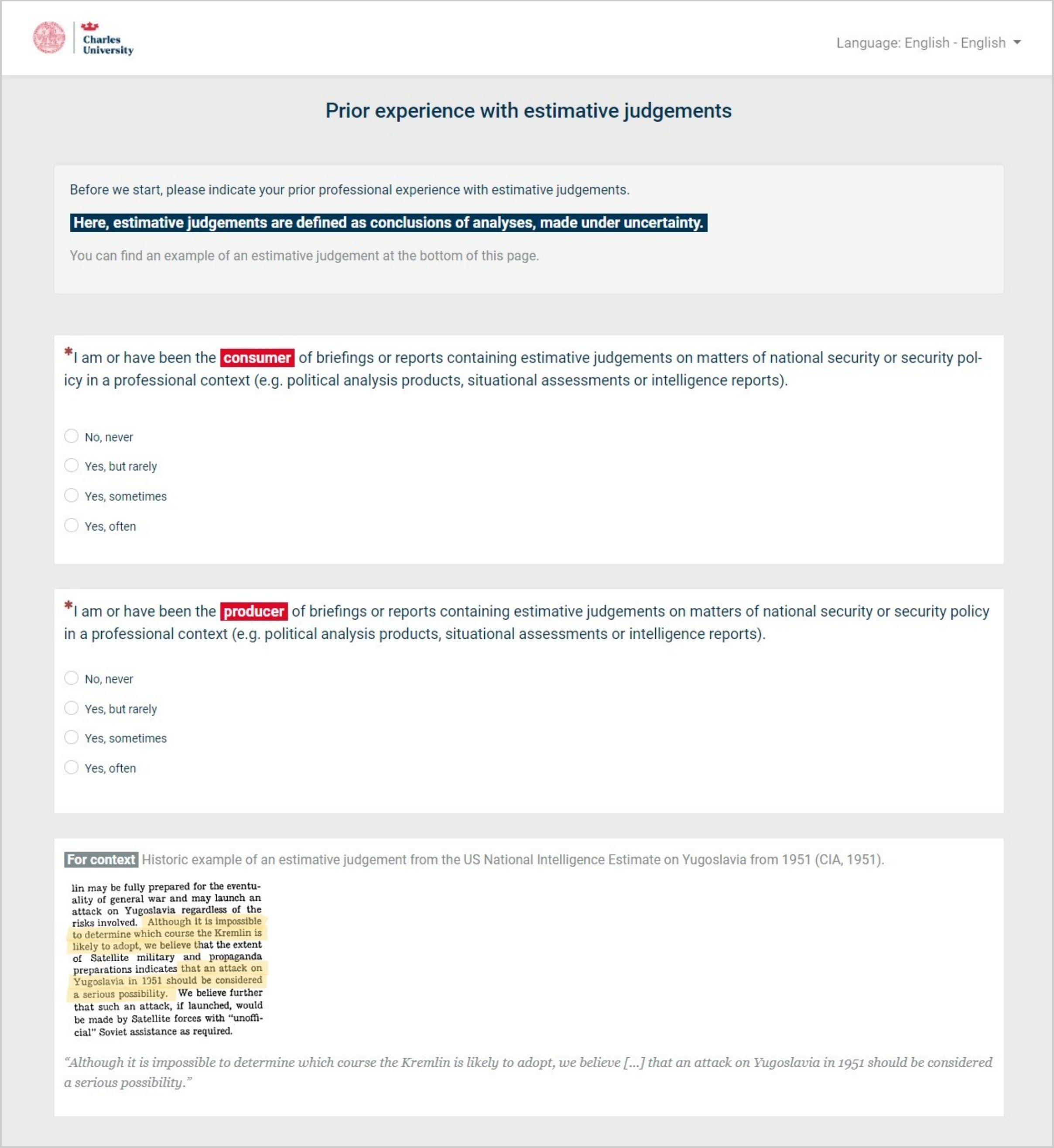
Figure 3.3: *Survey page (excerpt) asking about the participant's own expertise*

The second part of the self-evaluation phase is concerned with the participants' prior professional experience in producing and/or consuming estimative judgements on matters of national and international security. For both the production and for the consumption, participants can select one of the following options:

1. No, never
2. Yes, but rarely
3. Yes, sometimes
4. Yes, often

[22] Subjects that the participants have "advanced knowledge" (level 3) of are preferred; "fundamental knowledge" (level 2) and "excellent knowledge" (level 4) rank second and third; subjects in which the participant either has "rudimentary awareness" (level 1) or "recognised expert knowledge" (level 5) are avoided as best as possible. A detailed outline of the selection algorithm is provided in Appendix no. 2.

To illustrate the idea of an estimative judgement to the participants, a historic example from the 1951 US NIE on Yugoslavia is also provided (Fig. 3.4).



Charles University

Language: English - English

Prior experience with estimative judgements

Before we start, please indicate your prior professional experience with estimative judgements.

Here, estimative judgements are defined as conclusions of analyses, made under uncertainty.

You can find an example of an estimative judgement at the bottom of this page.

*I am or have been the **consumer** of briefings or reports containing estimative judgements on matters of national security or security policy in a professional context (e.g. political analysis products, situational assessments or intelligence reports).

○ No, never
○ Yes, but rarely
○ Yes, sometimes
○ Yes, often

*I am or have been the **producer** of briefings or reports containing estimative judgements on matters of national security or security policy in a professional context (e.g. political analysis products, situational assessments or intelligence reports).

○ No, never
○ Yes, but rarely
○ Yes, sometimes
○ Yes, often

For context Historic example of an estimative judgement from the US National Intelligence Estimate on Yugoslavia from 1951 (CIA, 1951).

lin may be fully prepared for the eventuality of general war and may launch an attack on Yugoslavia regardless of the risks involved. Although it is impossible to determine which course the Kremlin is likely to adopt, we believe that the extent of Satellite military and propaganda preparations indicates that an attack on Yugoslavia in 1951 should be considered a serious possibility. We believe further that such an attack, if launched, would be made by Satellite forces with "unofficial" Soviet assistance as required.

"Although it is impossible to determine which course the Kremlin is likely to adopt, we believe [...] that an attack on Yugoslavia in 1951 should be considered a serious possibility."

Figure 3.4: *Survey page asking about the participant's prior professional experience*

The self-evaluation of the participants' prior professional experience factors into the later distinction between different cohorts of 'expert participants' and 'non-expert participants'.

Phase II – Estimate Production

The second phase consists of two tasks for the participant: Estimate creation and estimate submission. Participants have to create and submit estimates on three scenarios that correspond to the three subject areas asked about previously:

1. “Poland leaves NATO within the next ten years.”
2. “Crimea remains under Russian control from 2024 to 2026.”
3. “China attacks Taiwan by the middle of this century.”

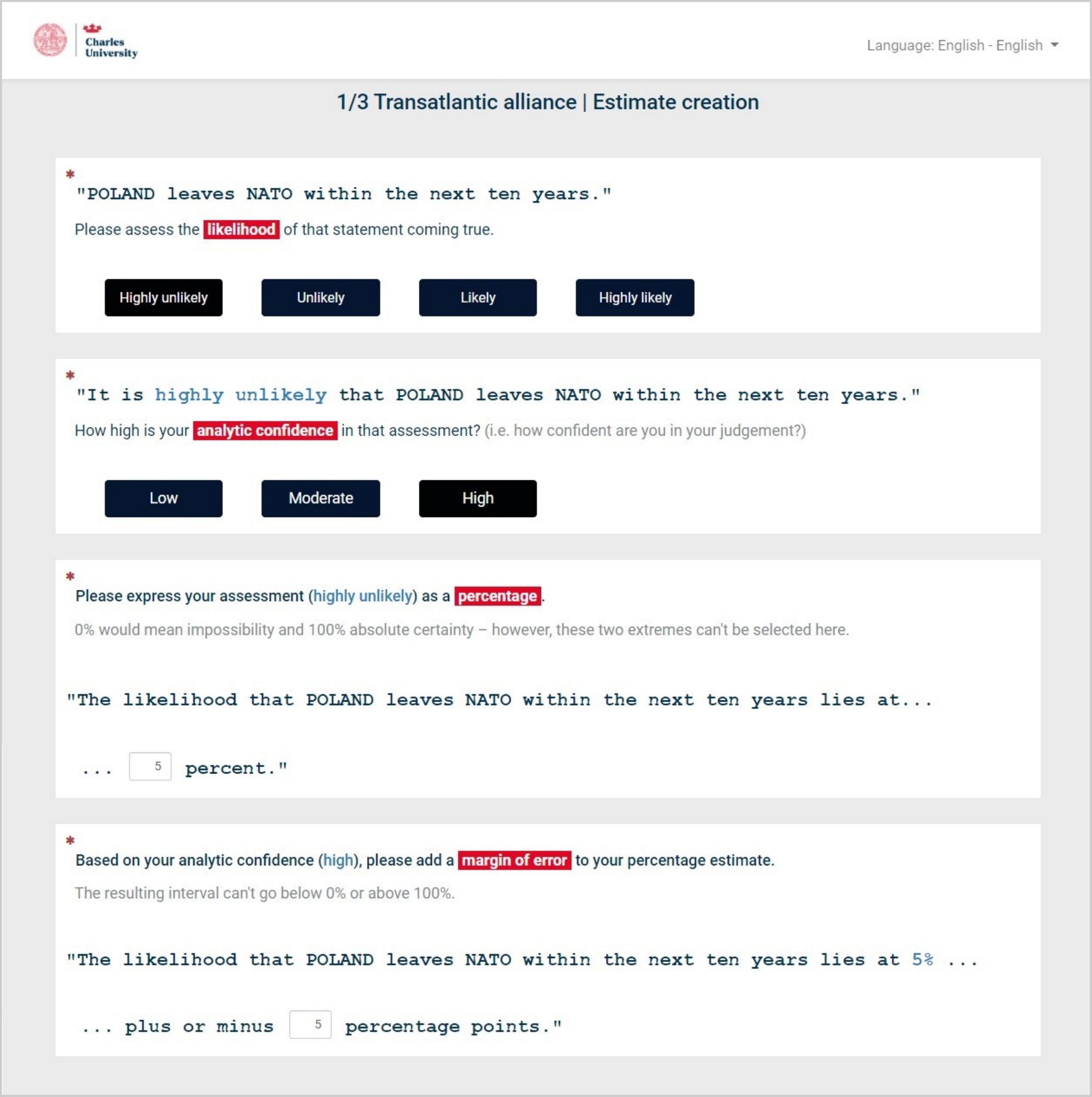
For each of the three statements, participants first have to assess the likelihood of that scenario as a verbal probability expression (VPE). They can choose between four options:

highly unlikely	unlikely	likely	highly likely

Afterwards, participants are asked to express their their level of analytic confidence in this assessment on a three-tiered ordinal scale:

low	moderate	high

Then, the participants have to translate their verbal assessments into numerical formats. First, they are asked to express their VPE assessment on the likelihood as a numerical point estimate (in percent). Finally, they have to add a margin of error to that point estimate (in percentage points), based on their level of analytic confidence (Fig. 3.5).



1/3 Transatlantic alliance | Estimate creation

* "POLAND leaves NATO within the next ten years."

Please assess the likelihood of that statement coming true.

* "It is highly unlikely that POLAND leaves NATO within the next ten years."

How high is your analytic confidence in that assessment? (i.e. how confident are you in your judgement?)

* Please express your assessment (highly unlikely) as a percentage.

0% would mean impossibility and 100% absolute certainty – however, these two extremes can't be selected here.

"The likelihood that POLAND leaves NATO within the next ten years lies at...

... 5 percent."

* Based on your analytic confidence (high), please add a margin of error to your percentage estimate.

The resulting interval can't go below 0% or above 100%.

"The likelihood that POLAND leaves NATO within the next ten years lies at 5% ...

... plus or minus 5 percentage points."

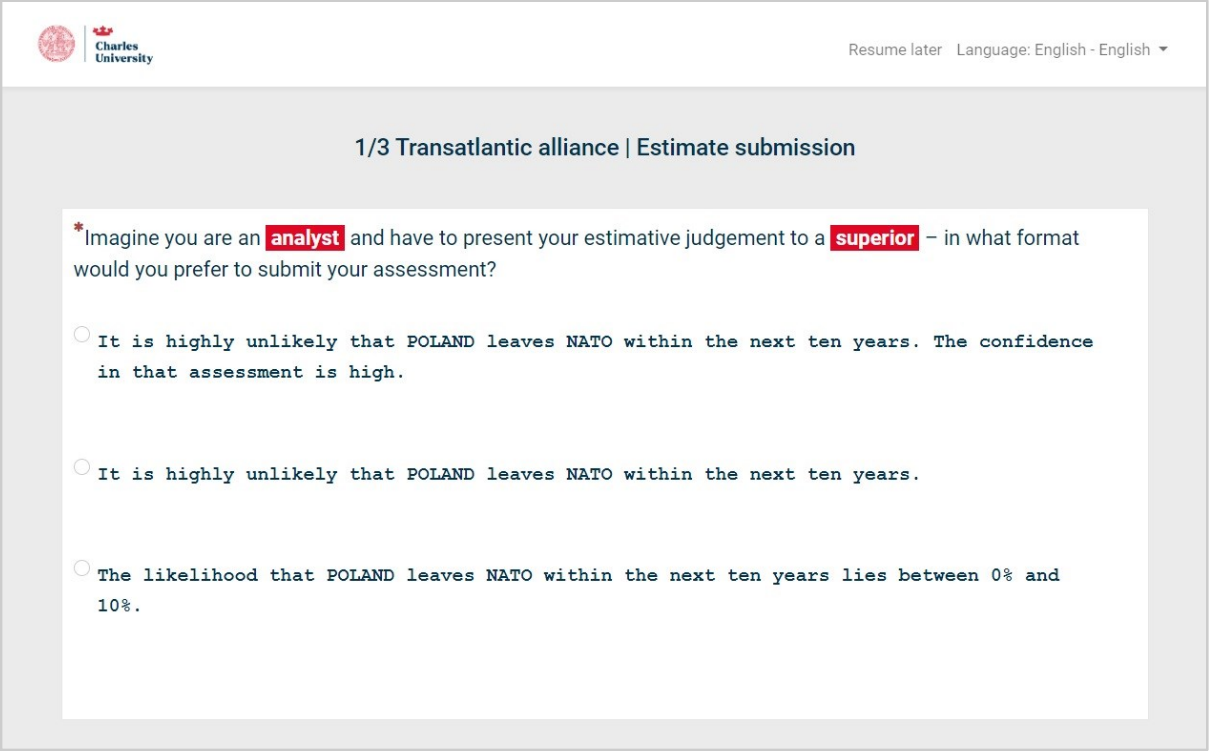
Figure 3.5: *Survey page for the estimate creation on the first subject (transatlantic alliance)*

The scenarios to be assessed for each subject area are deliberately chosen in order to yield different levels and combinations of likelihood and confidence. Poland leaving NATO is assumed to produce low likelihoods with high levels of analytic confidence; Russia staying in control over Crimea is expected to be assessed as rather likely with moderate to high levels of analytic confidence; for China attacking Taiwan, the estimates are expected to be pretty widely spread across the range with rather low levels of analytic confidence.

Each estimate creation is followed by the estimate submission. In that part of the survey, participants have to select in which format they would submit their previously made assessment to a superior. Participants can choose between:

1. Their assessment of the likelihood in the form of the VPE they selected;
2. Their VPE-assessment of the likelihood, but with the addition of a verbal statement on their level of analytic confidence;
3. A numerical range estimate, created based on their point estimate and the margin of error.

The three format options are presented in a random order (Fig. 3.6).



1/3 Transatlantic alliance | Estimate submission

*Imagine you are an **analyst** and have to present your estimative judgement to a **superior** – in what format would you prefer to submit your assessment?

○ It is highly unlikely that POLAND leaves NATO within the next ten years. The confidence in that assessment is high.

○ It is highly unlikely that POLAND leaves NATO within the next ten years.

○ The likelihood that POLAND leaves NATO within the next ten years lies between 0% and 10%.

Figure 3.6: *Survey page for the estimate submission on the first subject (transatlantic alliance); inputs were those shown in Fig. 3.5*

The process of estimate creation and estimate submission is repeated for each of the three subject areas.

Intermission

Between phase two and three, participants are asked to provide some personal information: Age (in five year-brackets), country of origin (either the nationality or the country where the participant has been living/working in for at least the last five years) and current occupation. Depending on the answer to the last question, the questionnaire is dynamically amended to refine the answer (Table 3.1).

1st layer	2nd layer		3rd layer	
Student or academic	Academic field	10 options + “other” (free text input)		
Public service (excluding academics)	Branch of public service	Ministry	Department	3 options + “other” (free text input)
		Government agency		4 options + “other” (free text input)
		Armed forces		
		Other (free text input)		
Political party or party-affiliated organisation	Kind of party-affiliated work	Elected representative		
		Party official		
		Staffer of an elected representative or party		
		Party-affiliated foundation		
Private sector	Main field of work	Consulting		
		Analysis (e.g. OSINT)		
		Defence industry		
		Other (free text input)		
Other (free text input)				

***Table 3.1**: Outline of the three layers of questioning on the participant’s current occupation*

Phase III – Review of other Participants' Estimates

In phase four, the participant is presented with the assessments of five other participants on one of the three subject areas. The other participants' assessments are displayed in whichever format they each chose to submit their estimate. That is, at least, what the participant is told – in reality, the five assessments are all automatically generated based on the participant's own estimates. The following variations are presented to the participant (in a random order):

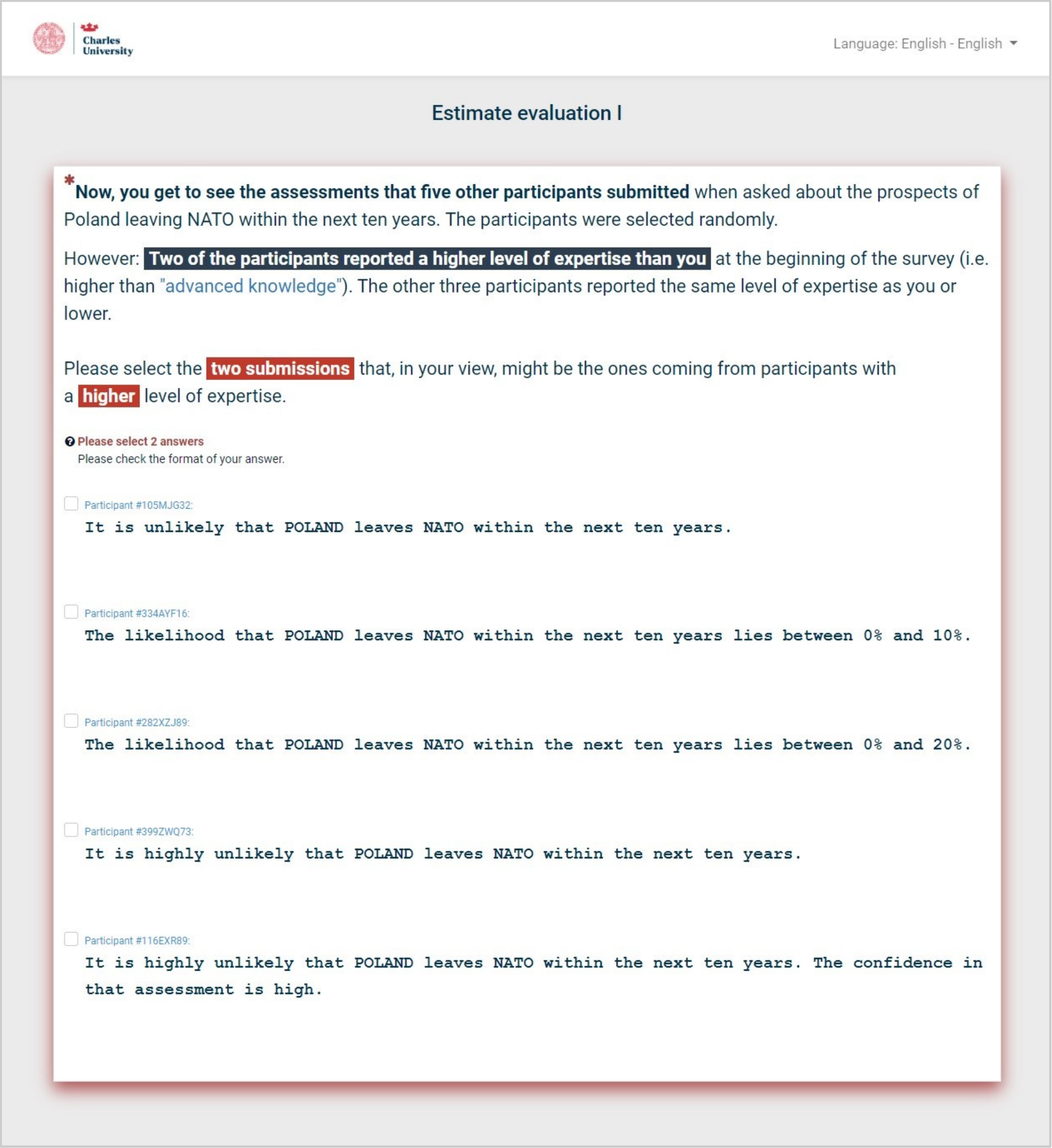
1. The same VPE on likelihood as the participant selected ("$\text{VPE}_{\text{original}}$")
2. The same VPE paired with the same level of analytic confidence as the participant originally assessed ("$\text{VPE+C}_{\text{original}}$")
3. The same numeric range estimate as the participant assessed ("$\text{NUM}_{\text{original}}$")
4. A VPE on likelihood that differs slightly[23] from the participant's original assessment ("$\text{VPE}_{\text{altered}}$")
5. A numeric range estimate that is slightly more conservative[24] than the participant's original range estimate ("$\text{NUM}_{\text{altered}}$")

Given these five assessments – that do not substantially differ in substance (from each other and from the participant's own beliefs) but only in format –, the participant has to select the two assessments that were supposedly made by other participants that reported a higher expertise in the given subject matter at the beginning of the survey (Fig. 3.7).[25]

[23] For the Poland/NATO and the Crimea/Russia scenarios: If the participant chose "highly likely" or "highly unlikely", the altered estimate would say "likely" or "unlikely". The same logic would apply in the opposite direction ("un-/likely" would become "highly un-/likely"). In the case of the Taiwan/China scenario, "highly un-/likely" in the original assessment would also become "un-/likely" in the altered version. However, if the original assessment said "likely", the altered version would say "unlikely" and vice versa ("unlikely" would become "likely").

[24] The newly generated range estimate would always contain the original range estimate of the participant. However, it would be wider with a tendency towards the middle of the range (50%).

[25] If, in phase one, a participant answered to have the highest level of expertise ("recognised expert knowledge") in every of the three subject areas, phase three would be skipped altogether.



Estimate evaluation I

***Now, you get to see the assessments that five other participants submitted** when asked about the prospects of Poland leaving NATO within the next ten years. The participants were selected randomly.

However: **Two of the participants reported a higher level of expertise than you** at the beginning of the survey (i.e. higher than "advanced knowledge"). The other three participants reported the same level of expertise as you or lower.

Please select the **two submissions** that, in your view, might be the ones coming from participants with a **higher** level of expertise.

Please select 2 answers
Please check the format of your answer.

□ Participant #105MJG32:
It is unlikely that POLAND leaves NATO within the next ten years.

□ Participant #334AYF16:
The likelihood that POLAND leaves NATO within the next ten years lies between 0% and 10%.

□ Participant #282XZJ89:
The likelihood that POLAND leaves NATO within the next ten years lies between 0% and 20%.

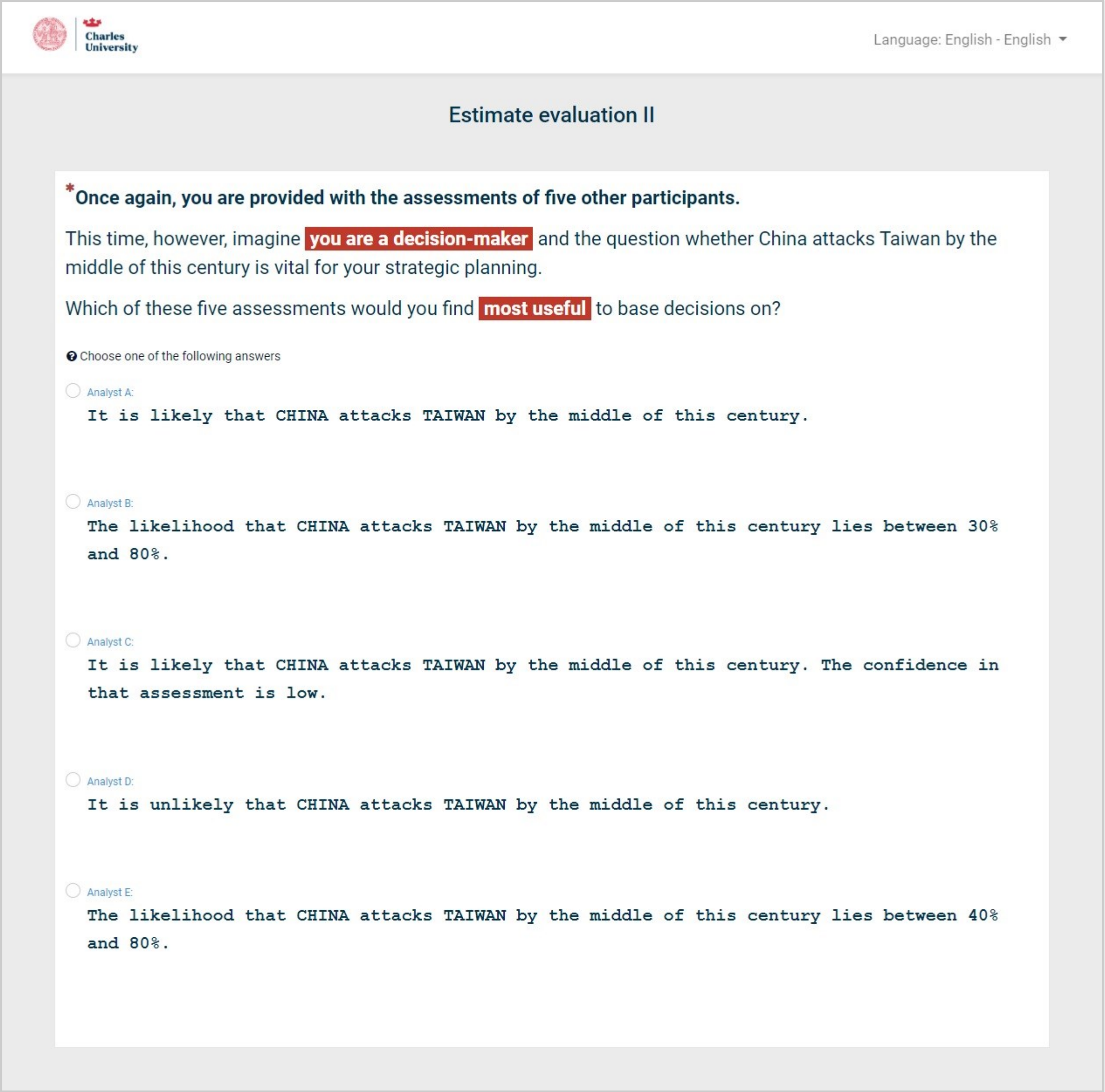
□ Participant #399ZWQ73:
It is highly unlikely that POLAND leaves NATO within the next ten years.

□ Participant #116EXR89:
It is highly unlikely that POLAND leaves NATO within the next ten years. The confidence in that assessment is high.

Figure 3.7: *Survey page for the selection of the two 'expert-assessments'. Assessments are presented in the following order: 1. VPE$_{altered}$; 2. NUM$_{original}$; 3. NUM$_{altered}$; 4. VPE$_{original}$; 5. VPE+C$_{original}$. The inputs from the original assessment were "highly unlikely", "high" analytic confidence, a point estimate of 5% and a margin of error of 5%; the given level of expertise was "advanced knowledge" (level 3).*

Phase IV – Consumer Format Preference

Phase four again elicits the format preferences of the participants – as in phase 2 – but from a consumer perspective. Participants are asked to choose one assessment out of five, which they would find most useful as a basis for decision-making (Fig. 3.8).



Estimate evaluation II

*Once again, you are provided with the assessments of five other participants.

This time, however, imagine **you are a decision-maker** and the question whether China attacks Taiwan by the middle of this century is vital for your strategic planning.

Which of these five assessments would you find **most useful** to base decisions on?

Choose one of the following answers

○ Analyst A:
It is likely that CHINA attacks TAIWAN by the middle of this century.

○ Analyst B:
The likelihood that CHINA attacks TAIWAN by the middle of this century lies between 30% and 80%.

○ Analyst C:
It is likely that CHINA attacks TAIWAN by the middle of this century. The confidence in that assessment is low.

○ Analyst D:
It is unlikely that CHINA attacks TAIWAN by the middle of this century.

○ Analyst E:
The likelihood that CHINA attacks TAIWAN by the middle of this century lies between 40% and 80%.

Figure 3.8: *Survey page for the consumer format preference. Here, the assessments are presented in the following order: 1. VPE$_{original}$; 2. NUM$_{altered}$; 3. VPE+C$_{original}$; 4. VPE$_{altered}$; 5. NUM$_{original}$. The inputs from the original assessment were "likely", "low" analytic confidence, a point estimate of 60% and a margin of error of 20%.*

Again, the participants are told that the five assessments presented came from other participants' submissions. As in phase three, this is not really the case; instead, the assessments are again generated based on the participant's own estimates.[26] Additionally, the participants are also asked about the reasons for their choice of preferred format. They can select one or more options out of the following:

[26] The algorithm to generate the assessments based on the participant's original inputs is the same as in phase three. However, the survey is designed in such a way that the scenario used in phase four is always different from the scenario presented in phase three.

1.1 It aligns with my own assessment

1.2 It differs from my own assessment

2.1 It is precise

2.2 It leaves room for ambiguity / It is not overly-specific

3. It is easily comprehensible

4. It treats uncertainties transparently

5. It contains a high analytic confidence

6. Other (free text input)

Conditional last Question – Real-World Format Usage

If a participant answers to be the producer of estimative judgements on national and international security in a professional capacity at the beginning of the survey, the survey is extended by one final question. Participants are asked to select all the formats in which they convey estimative judgements in their professional life. The provided options are:

1. Non-standardised verbal probability expressions
   (e.g. "probably", "doubtful" etc.)
2. Standardised words of estimative probability
   (e.g. according to NATO guidelines)
3. Standardised or non-standardised words of estimative probability
   with a statement on analytic confidence (e.g. high/moderate/low)
4. Point estimates (e.g. “75% likelihood”)
5. Numeric ranges (e.g. “likelihood between 70% and 80%”)
6. Other (free text input)

3.4 Analytical Aims

The survey design's principal aim is to address the three main research hypotheses laid out previously (see chapter *3.1 Research Hypotheses*) – they constitute the primary analytical aims of this project. However, the survey also produces data on a wide range of additional aspects. These are the secondary analytical aims of this research.

Hypothesis 1 is addressed in phase two and four of the survey. Phase two puts participants in the position of an estimate producer and elicits their format preferences in that role. The submission of the estimate is deliberately framed in such a way that the participant feels accountable vis-à-vis a superior. In phase four, on the other hand, participants are not personally accountable. The framing of this part of the survey puts no hierarchical pressure on the participants and they bear no responsibility for the substance of the assessments they are presented with (since they are told that these assessments came from other participants). Instead, participants are asked about their format preferences purely on the grounds of the assessments' usefulness as a basis for decision-making.

Phase two also provides the data to test ***hypothesis 2*** – the congruence principle. As mentioned previously, the scenarios to be assessed are deliberately chosen in order to produce estimates with varying degrees of certainty. Particularly, the estimate on Poland leaving NATO within the next ten years is expected to yield estimates with a high degree of certainty by the participants – (very) low likelihood estimates with high levels of analytic confidence. The estimates on China attacking Taiwan until 2050, on the other hand, are expected to feature much higher levels of uncertainty. The data obtained from phase two of the survey is intended to provide insights into whether format preference is sensitive to this difference in certainty.

The last primary analytical aim of the survey is to test ***hypothesis 3*** – the illusion of rigour. The data needed will be produced in phase three of the survey. Participants are presented with an array of assessments that do not differ much in substance. They also closely reflect the participant's own assessment on the subject matter at hand. Nevertheless, participants are forced to select two assessments that, in their eyes, might be the ones made by participants with a higher self-reported expertise. The assessments are also listed in a random order. This task is designed to minimise any factors that might potentially influence the decision-making on which submissions are the 'expert-assessments' – except for the differences in format to communicate the assessments. If format is not a factor in this context, the selection should be random, meaning that the choice of assessments deemed to be 'expert-made' should be distributed pretty evenly across all options provided. This null hypothesis will be rejected if numerical formats are significantly more often chosen as the 'expert-assessments', suggesting a potential illusion of rigour.

The first among the ***secondary research aims*** is comparing the responses from experts and non-experts. Experts in this context are persons who professionally produce or consume intelligence products or other kinds of analyses on national and international security. On the consumer-side, these are, for instance, high-level decision makers working on security policy in the legislative and executive branches of government. On the producer-side, intelligence analysts constitute the prime expert-cohort. Mainly due to the high profile (in the case of decision-makers) and high levels of confidentiality (in the case of the intelligence community), these expert-subjects are very difficult to access. That's why this research, similar to past research in that field (e.g. Friedman et al., 2017; Irwin/Mandel, 2023), combines an expert cohort with a non-expert cohort. The non-expert cohort mainly serves the purpose of generating a dataset large enough to yield statistically significant results. The expert cohort, on the other hand, has a control function: The results from the larger non-expert cohort are only indicative as long as non-expert responses do not differ too much from expert responses.

Besides the prior professional experience with estimative judgements, the survey also elicits the self-ascribed expertise in the three subject areas that the participants have to make assessments on. That will allow the analysis to also explore possible relationships between (self-ascribed) expertise and the substance of the estimates; for instance, whether a higher level of expertise correlates with more conservative or cautious estimates.[27]

Lastly, the experiment also produces further data on the numeric interpretation of VPEs and verbal levels of confidence. Although there already exists research on that topic, the additional data that this survey produces is an unavoidable by-product which might, nevertheless, be interesting to analyse.

3.5 Limitations

There are some limitations that this project faces – some inherent, some caused by deliberate decisions in the conceptualisation and design of the project. These limitations can be categorised into methodological limitations and practical limitations.

Methodological Limitations

Firstly, the methodological framework as well as the existing literature treats verbal and numerical formats as general categories. In the survey, however, not every possible verbal and numerical format can be featured (see chapter *3.2 Selection of tested formats*). In the methodology and analysis, the chosen formats are somewhat treated as representative of the broader category of formats they belong to (i.e.

[27] That would mean lower levels of analytic confidence, larger margins of error and potentially less extreme likelihood estimates (i.e. likelihood estimates that are further away from the extreme ends 0% and 100%).

‘verbal formats’ versus ‘numerical formats’). But one should remain conscious of the fact that different variants of verbal and numerical formats might yield slightly different results. Generalising the results from the survey to make statements about the entirety of verbal and numerical formats should, therefore, be approached with caution.

Secondly, with regards to the format preferences, the research can only investigate whether the divergence between estimate producers and consumers described in the literature is reproducible in an experimental setting. If that were the case, it would demonstrate that one and the same person, when presented with one and the same assessment, has differing preferences regarding the format – depending on their communicative role (either at the producing or receiving end of the estimative judgement). That does not, however, explain the degree to which this change in preference is caused by the face-management function of verbal vagueness. That said, it is generally quite challenging to elicit format preferences without face-management somehow influencing the results. Where there is communication of any kind, there is a sender and a receiver (of an estimative judgement for example). It is quite reasonable to assume that this sender-receiver dynamic always triggers face-management considerations on the sender’s part which influence format preference.

A third methodological limitation is introduced through the participants’ self-evaluation regarding their prior experience with estimative judgements. At the beginning of the survey, participants are asked whether they produce estimative judgements on matters of national security in a professional capacity. The same is also asked about the consumption of such analysis products. To answer, participants can choose between “No, never” and three ascending degrees of “Yes” (from “rarely” to “often”). Except for the option “No, never”, this is a highly subjective measure, since there are no definitions on what exactly qualifies as “rarely”, “sometimes” or “often”. That makes a comparison between participants difficult – for the same degree of prior experience, one participant might answer “sometimes” while another would already call that “often”. This limitation has to be considered when discriminating non-experts from expert participants. For that reason, the analysis will consider further criteria to define the ‘expert’-cohort.

The self-evaluation of the participant’s knowledge in the three subject areas (also in phase one of the survey) encounters the same limitations as the self-evaluation of prior experience. But since the assessment on the participant’s knowledge mainly serves the purpose to determine the later course of the survey and is not aimed at yielding an objective measure of the participants’ expertise, this limitation is not as grave.

Practical Limitations

The biggest practical limitation is getting access to top-tier ‘expert’-participants. Due to the sensitive nature of the subject matter (i.e. intelligence reports and other analyses on national security) it is a serious challenge for this project to get real-world analysts and decision-makers to participate in the survey. How the research deals with that challenge is outlined in the following section on data collection.

A second practical limitation is introduced through the fact that the survey has to work in three languages – English, German and French. This, in turn, is a product of the aforementioned challenge of getting access to expert-participants. Making the survey available in multiple languages increases the number of potential participants – which is deemed necessary, given the challenge of expert-access. That decision, however, poses restrictions on the survey design, as addressed in chapter *3.2 Selection of tested formats*.

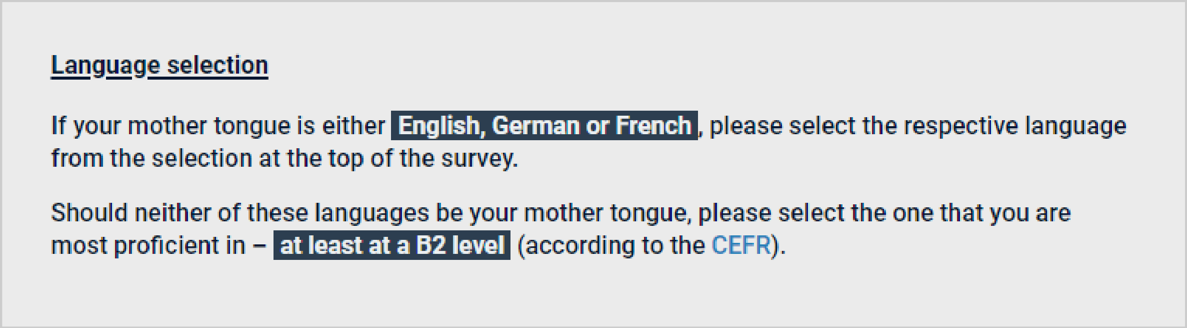
A third practical limitation is encountered in phase three and four as well as in the subsequent analysis of the data from those two phases. As laid out in the survey design, each participant will be presented twice with supposed assessments of other participants. In reality, the assessments are all generated based on the participants own assessments in order to mitigate any possible influence of a difference in substance on the participant's preferences and perceptions. For this 'illusion' to work, however, the presented assessments cannot *all* be identical to the participant's original estimates.[28] Therefore, in addition to the participant's original estimate in the three formats featured in the research, also two 'altered' estimates are added. That complicates the data analysis significantly. Without the altered estimates, each of the three formats would have had an equal baseline-probability of being selected by the participant of $\frac{1}{3}$. Any statistically significant deviation from that equal distribution in favour of one format would have indicated an effect on perception or preference. With the introduction of the two altered assessments the analysis is not as simple anymore. It is believed, however, that the inclusion of two altered estimates is necessary to 'sell the illusion'.

3.6 Data Collection

Data was collected over the span of one month from 4 June to 4 July 2024. The survey was created in and hosted on LimeSurvey, an open source online statistical survey web app. To access the survey, participants had to click on a link that was sent to them; no access code or token was necessary. The survey did, however, use cookies to prevent repeated participation. The survey was available in three languages: English, German and French. Before the start of the survey, participants

[28] The fact that, within the presented assessments, each of the three formats has to be represented at least once, further exacerbates the challenge of making the mock-participants' estimates believable. Just taking the participant's original estimates and presenting it in the three formats once each will certainly not look like a random sample of other participants' submissions, as the survey claims.

were instructed to select the language they were most proficient in and in which they had at least a B2 (upper-intermediate) level (Fig 3.9). That was to ensure that there would be no language barrier influencing the participants' preferences and perceptions later on in the survey.



Language selection

If your mother tongue is either English, German or French, please select the respective language from the selection at the top of the survey.

Should neither of these languages be your mother tongue, please select the one that you are most proficient in – at least at a B2 level (according to the CEFR).

Figure 3.9: *Excerpt of the landing page before the start of the survey*

For the non-expert cohort, students constituted the main target group. Participation in the survey was not limited to students of a specific faculty, university, or even country. However, most accessible students for this experience have a background in social sciences. In addition to the student cohort, also 'experts' were contacted to take part in the survey. 'Experts' were defined as people who professionally produce and/or consume analyses on national and international security – in the intelligence community, in ministries, in the armed forces, in the legislative, in international organisations, in the academic world, in think tanks or in the private sector.

4. Data Analysis

4.1 Descriptive Statistics

Over the course of one month of data collection, 153 participants completed the survey. Additionally, 59 participants took part in the survey but did not finish it. These incomplete responses will not be considered in the data analysis.

Subject Attributes

Most participants completed the survey in German (51%) or English (43.1%) with only a small share of respondents opting for the French version (5.9%). In terms of nationality[29], German nationals account for half (50%) of the dataset.

German	77 (50.7%)	Ukrainian	3 (2.0%)	Indian	1 (0.7%)
Italian	10 (6.6%)	Armenian	2 (1.3%)	Iraqi	1 (0.7%)
Czech	9 (5.9%)	Polish	2 (1.3%)	Norwegian	1 (0.7%)
French	9 (5.9%)	Portuguese	2 (1.3%)	Russian	1 (0.7%)
Belgian	5 (3.3%)	Spanish	2 (1.3%)	Slovakian	1 (0.7%)
British	4 (2.6%)	Tunisian	2 (1.3%)	Slovenian	1 (0.7%)
American	3 (2.0%)	Austrian	1 (0.7%)	Swedish	1 (0.7%)
Canadian	3 (2.0%)	Bosnian	1 (0.7%)	Turkish	1 (0.7%)
Dutch	3 (2.0%)	Greek	1 (0.7%)	No answer	6 (3.9%)

Table 4.1: *Participants' nationalities; participants per nationality in absolute numbers, share of the total dataset in brackets (N = 153)*

The other half of participants has a quite diverse national background. Out of the 25 non-German nationalities, no participant group has a share larger than 7% (Table 4.1). 6 participants (3.9%) chose not to answer that question.

[29] If participants live and/or work in a country different to their nationality for at least the last 5 years, they were asked to select the country they were living/working in at the moment.

Participants' age ranges from the 20-24 age bracket to the 65-69 age bracket (Fig. 4.1). The median age bracket is 25 to 29 years.

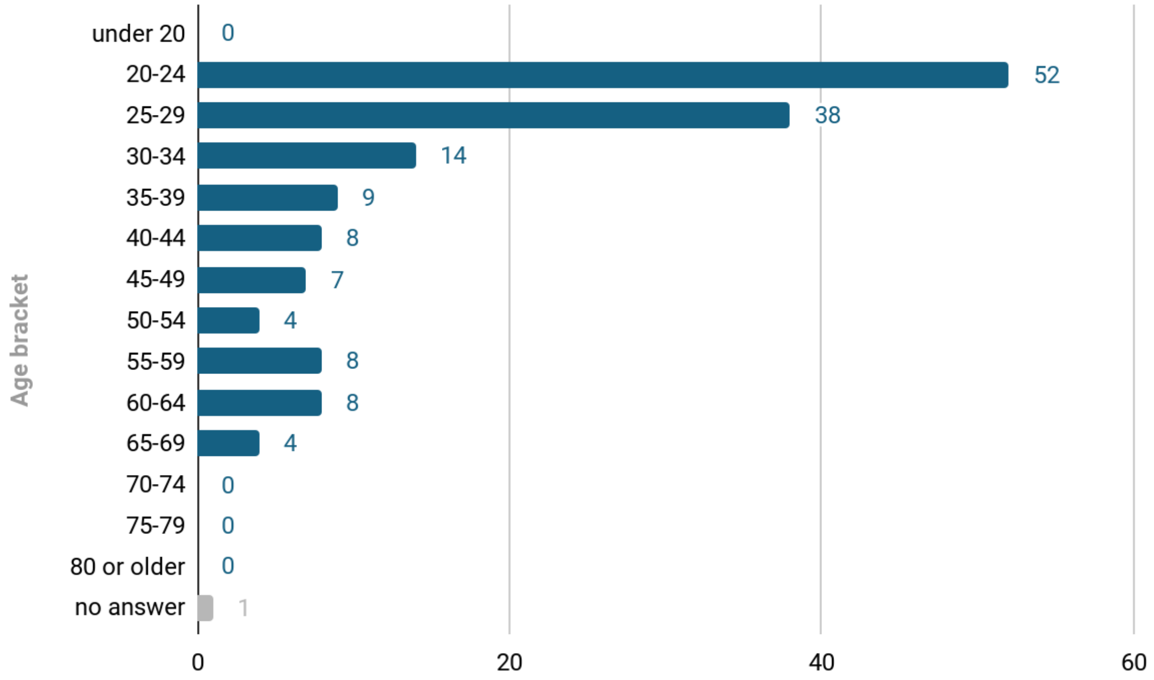


Figure 4.1: *Participants' age distribution (N = 153)*

This age distribution reflects the fact that students and young professionals were the most accessible demographic as participants (as discussed in *3.5 Limitations* and *3.6 Data Collection*).

About half of all participants (49.7%) reported to have no prior professional experience producing estimative judgements on matters of national security or security policy. Regarding the consumption of such assessments, only less than a quarter of respondents (23.5%) reported to have no prior professional experience (Fig 4.2). Going up the experience levels, the situation reverses: While 28.8% of participants report to often *consume* estimative judgements on security, only 10.5% of respondents often *produce* such analyses.

> *“I am or have been the consumer / producer of briefings or reports containing estimative judgements on matters of national security or security policy in a professional context (e.g. political analysis products, situational assessments or intelligence reports).”*

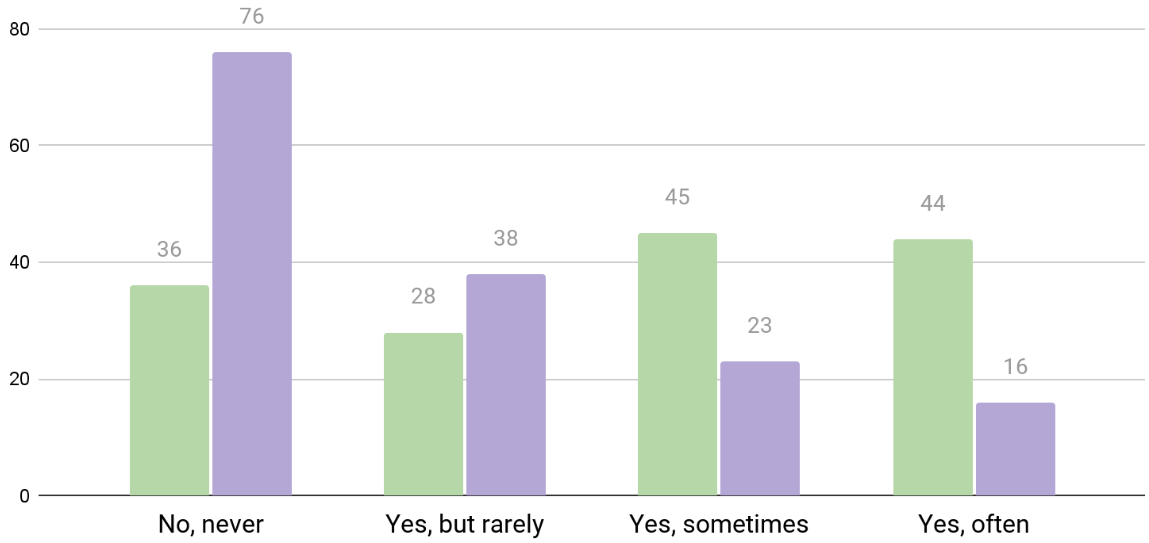


Figure 4.2: *Prior professional experience of participants as estimate producers and consumers (y-axis indicates number of respondents for each level of experience)*

The self-evaluation of participants’ prior professional experience is the starting point for defining an ‘expert’ cohort and a ‘non-expert’ cohort. This cumulative view of the data, however, is not sufficient to devise an adequate classification of cohorts.

Cohort Classification

Harkening back to the conceptual framework (*2.4 The Producer/Consumer Relationship*), the composition of experience levels for each participant has to be considered: A staff analyst would likely report a high level of experience for producing analyses but lower levels for consumption. A top-level decision-maker, on the other hand, does not have to produce analyses but might frequently consume them. Positions between those two extremes will have various ‘mixtures’ of consumption and production.

Table 4.2 maps all the absolute frequencies of all possible combinations of production and consumption levels that participants answered (the number in each cell reflects the number of participants with a given combination of experience levels). The aforementioned staff analyst would appear in or around the upper left corner of the matrix (high production, low consumption), whereas a high-level decision-maker would rank in or around the lower right corner (low production, high consumption). Someone in between those two extremes would be located in the upper right corner (high production, high consumption). Overall experience with estimative judgements in a security context decreases from the top right to top left.

consumer producer	No, never	Yes, but rarely	Yes, sometimes	Yes, often
Yes, often	1	0	1	14
Yes, sometimes	0	3	10	10
Yes, but rarely	2	3	17	16
No, never	33	22	17	4
Participant classification:	*No experience* **33** (21.6%)	*Low exp.* **27** (17.6%)	*Medium exp.* **47** (30.7%)	*High exp.* **46** (30.1%)

Table 4.2: *Professional experience as producers and consumers of estimative judgements; the higher the frequency, the darker cells are shaded*

This view of the data shows that, within this sample, the archetypical analyst is represented more than the archetypical decision-maker – numbers are higher on the lower right side of the matrix than towards the upper left side. This could either be a systematic sampling error due to accessibility limitations or, indeed, an accurate reflection of the true population since, arguably, in any given state there are more low-level analysts than top-level decision-makers.

For the purpose of this research, this data is only relevant for the delineation between ‘experts’ and ‘non-experts’: Based on the representation in figure 4.4, groups of different experience can be drawn to structure the dataset: No expertise, low expertise, medium expertise and high expertise (see bottom row in Table 4.2). But just taking the “high expertise” group and defining it as the ‘expert’ cohort would be premature.

A closer look at the data shows that the “high expertise” group still contains suspiciously many participants who are under 30 years of age and who answered “student or academia” as their current occupation. Presumably, some of the student participants studying Security Studies, International Relations or a related field overlooked the caveat that the question asked about *professional* experience – or, at least, interpreted that condition very liberally, counting university studies as professional experience. Whatever the exact reason, a further step to define the ‘expert’ cohort is needed: All respondents aged 29 or younger that *also* reported “academia” as their current occupation are omitted from this cohort and classified as non-experts. That leaves an expert cohort of 34 respondents.

The data suggests that this classification, indeed, is adequate. For instance, when looking at the self-ascribed expertise: At the beginning of the survey, participants were asked to rank their own level of expertise in three subject areas. Within the non-expert sample, the median self-evaluated level of expertise is “advanced knowledge” for the subject areas “NATO” and “Ukraine War”. Median self-ascribed expertise on “East Asian Security” is one level lower at “fundamental knowledge” for the non-experts.

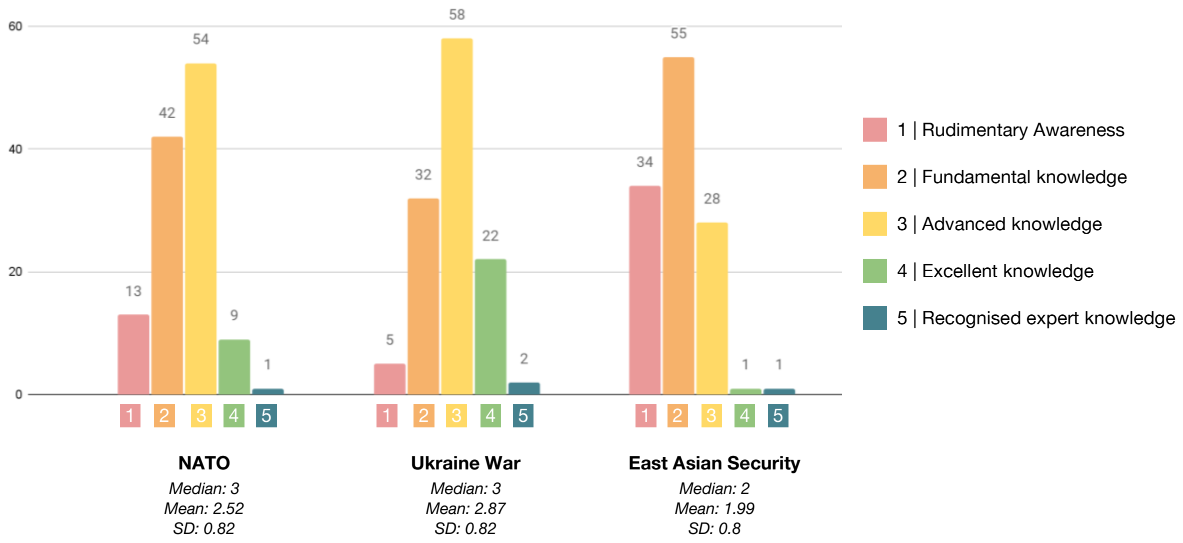


Figure 4.3.1: *Self-ascribed expertise in the* *<u>non-expert cohort</u>* *(n = 119)*

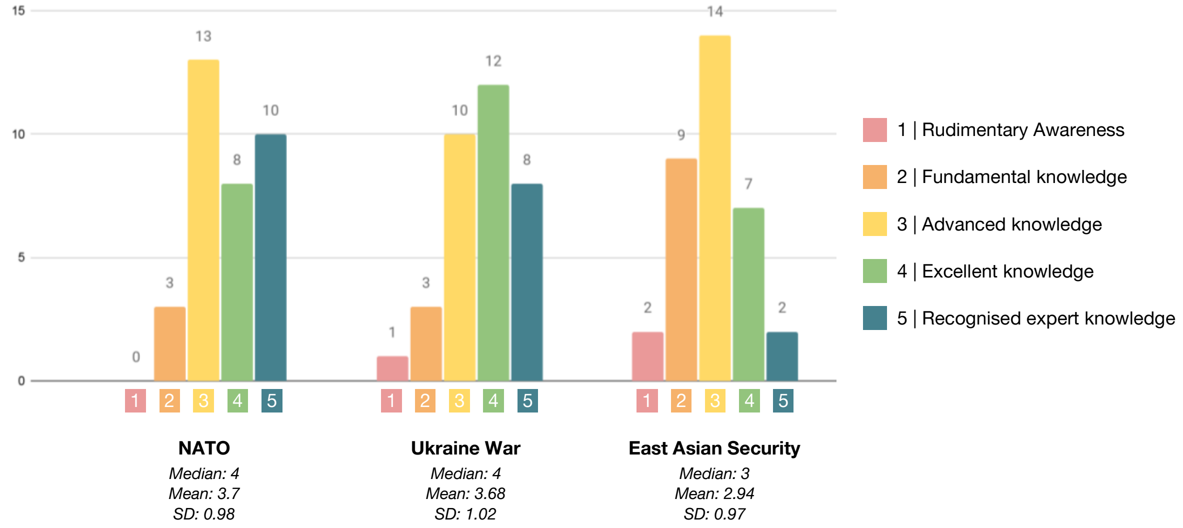


Figure 4.3.2: *Self-ascribed expertise in the* *<u>expert cohort</u>* *(n = 34)*

The expert cohort, on the other hand, reported higher levels of expertise for all three topics (compare Fig 4.3.1 and Fig 4.3.2). The differences in mean levels of expertise between the non-expert cohort and the expert cohort are statistically significant in all three cases (Table 4.3). That supports the assumption that participants with more experience also have a significantly higher expertise (albeit self-evaluated) in the three security-related subject areas that the survey asked about.

Expertise	NATO		Ukraine War		East Asian Security	
t-test (ass. uneq. Var.)	Non-Experts	Experts	Non-Experts	Experts	Non-Experts	Experts
Mean	2.521008	3.735294118	2.865546218	3.676470588	1.991596639	2.941176471
Variance	0.675402	0.988413547	0.676684233	1.073975045	0.635522005	0.966131907
Observations	119	34	119	34	119	34
df	47		46		46	
t Stat	-6.51425115		-4.200275698		-5.168385853	
P (T<=t) one-tail	< .001***		< .001***		< .001***	
P (T<=t) two-tail	< .001***		< .001***		< .001***	

Table 4.3: *Two-sample (non-experts vs. experts) t-test statistics for participants' mean expertise in each of the three subject areas covered in the survey*

The age distribution (Fig. 4.4) of the non-expert cohort still is heavily skewed towards participants in their twenties (median age bracket is also still 25-29) – for the reasons discussed earlier. The expert cohort, on the other hand, has a fairly even age distribution (and a median age of 45-49).

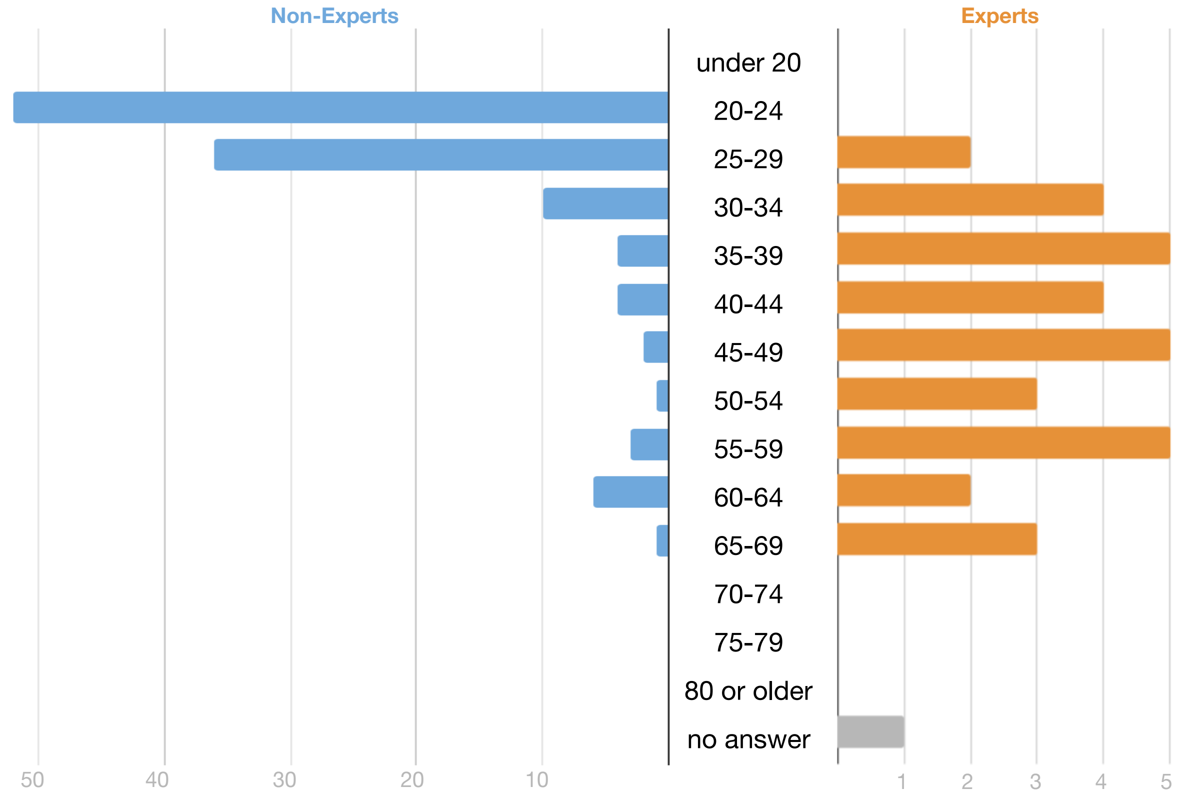
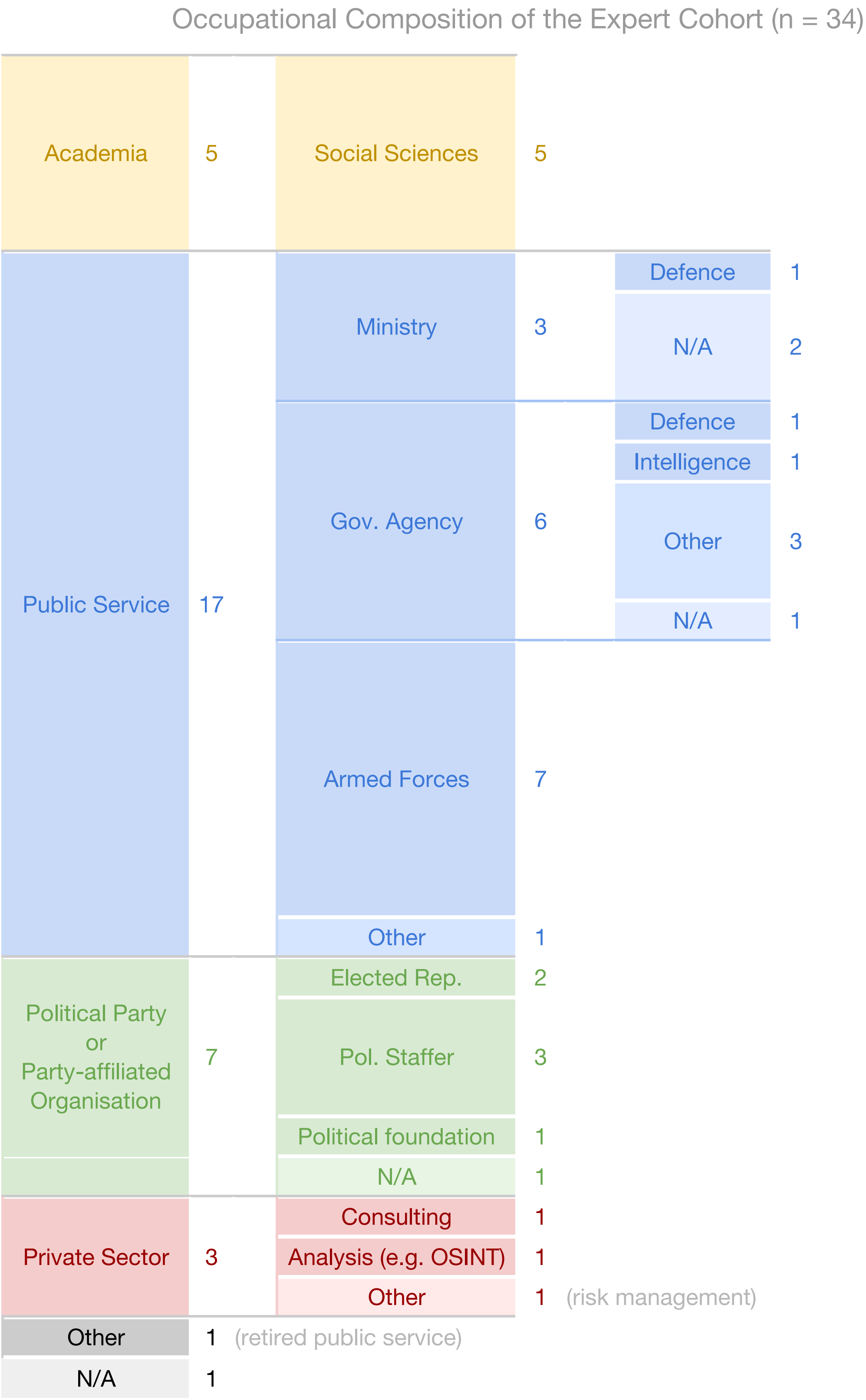


Figure 4.4: *Age distribution for the non-expert and expert cohort*



Occupational Composition of the Expert Cohort (n = 34)

Academia	5	Social Sciences	5			
Public Service	17	Ministry	3	Defence	1	
				N/A	2	
		Gov. Agency	6	Defence	1	
				Intelligence	1	
				Other	3	
				N/A	1	
		Armed Forces	7			
		Other	1			
Political Party or Party-affiliated Organisation	7	Elected Rep.	2			
		Pol. Staffer	3			
		Political foundation	1			
		N/A	1			
Private Sector	3	Consulting	1			
		Analysis (e.g. OSINT)	1			
		Other	1	(risk management)		
Other	1	(retired public service)				
N/A	1					

Figure 4.5: *Occupational information of the expert cohort, based on Table 3.1*

The vast majority of the expert cohort is German (79.4%) and expert participants' most common field of work is public service (50%). A detailed overview of expert participants' occupations (i.e. how many participants belong to each category) is provided in Fig. 4.5.

Submitted Assessments

In their verbal assessments of the three scenarios, experts and non-experts did not differ much. It is apparent, however, that the expert cohort has less outlier assessments.

For instance, the likelihood of Poland leaving NATO within the next ten years was assessed by all expert participants as "highly unlikely" or "unlikely" (Fig. 4.6.1). The majority of non-expert participants (93.2%) also responded "highly unlikely" or "unlikely", but a small fraction assessed "likely" or even "highly unlikely". A similar dynamic can also be observed in the answers for the Crimea/Russia scenario, but with "likely" and "highly likely" as the dominant assessments (Fig. 4.6.2). In their assessments on the likelihood of China attacking Taiwan, expert and non-expert assessments are very similar (Fig. 4.6.3).

Overall, the expert cohort reported slightly higher confidence levels. In each of the three scenarios, the share of experts that reported "high" confidence is greater than that of the non-experts – whereas the share of non-experts with a "moderate" confidence level is always greater than that of the experts (right side of Fig.4.6.1-Fig.4.6.3).

Poland leaves NATO within the next ten years.

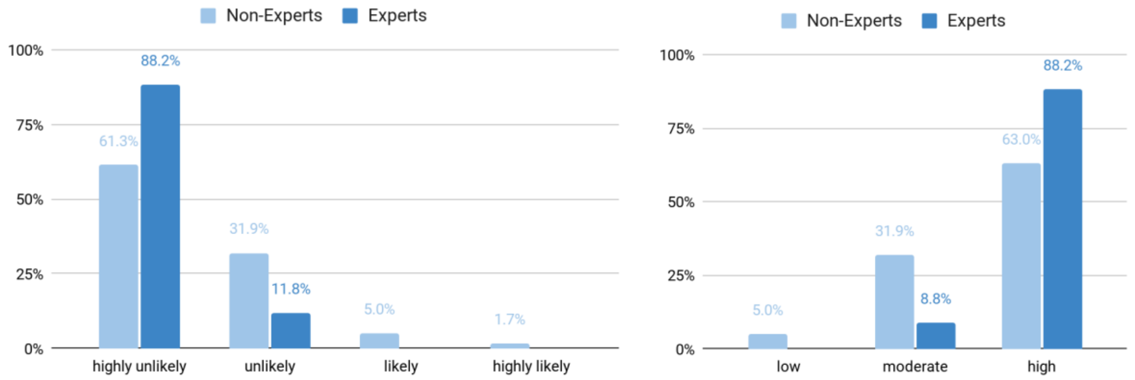


Figure 4.6.1: *Participants' verbal assessments of likelihood (left) and their analytic confidence (right)*

Crimea remains under Russian control from 2024 to 2026.

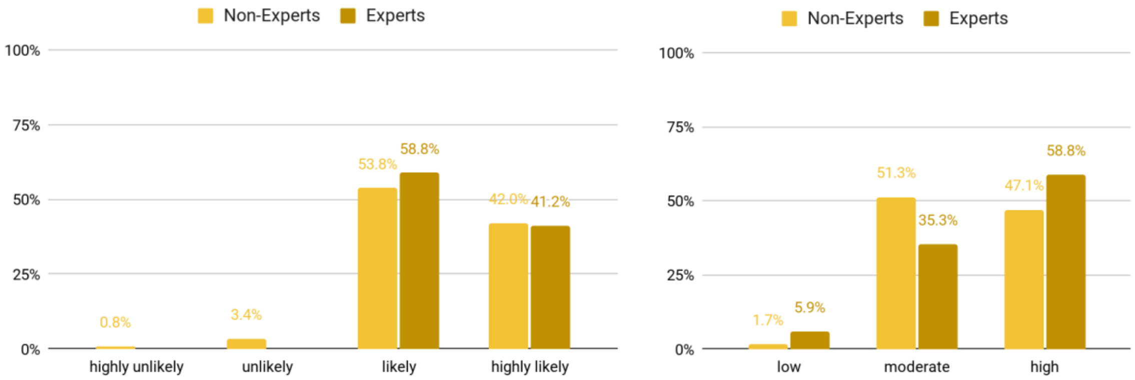


Figure 4.6.2: *Participants' verbal assessments of likelihood (left) and their analytic confidence (right)*

China attacks Taiwan by the middle of this century.

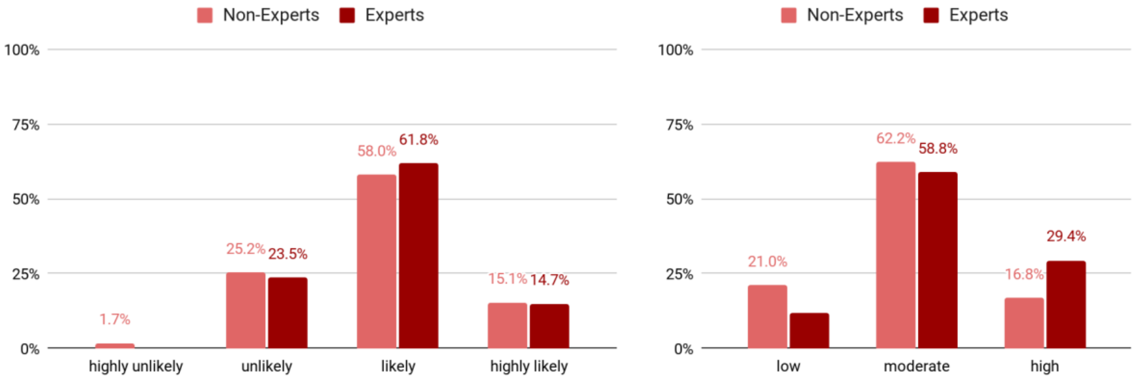


Figure 4.6.3: *Participants' verbal assessments of likelihood (left) and their analytic confidence (right)*

The numeric estimates broadly reflect the verbal estimates of each scenario (Fig. 4.7). Both likelihood estimates and error margins also do not differ significantly[30] between experts and non-experts (Tables 4.4.1 and 4.4.2).

Likelihood	NATO (Poland)		Ukraine War (Crimea)		East Asian Sec. (Taiwan)	
t-test (ass. uneq. Var.)	Non-Experts	Experts	Non-Experts	Experts	Non-Experts	Experts
Mean	23.86554622	18.29411765	74.91596639	78.61764706	54.77310924	61.23529412
Variance	863.3038029	984.1532977	356.2301666	146.4251337	427.5497792	554.1853832
Observations	119	34	119	34	119	34
df	51		84		48	
t Stat	0.925998425		-1.370039184		-1.448890315	
P (T<=t) one-tail	.179404055		.087162756		.076935023	
P (T<=t) two-tail	.358808109		.174325511		.153870047	

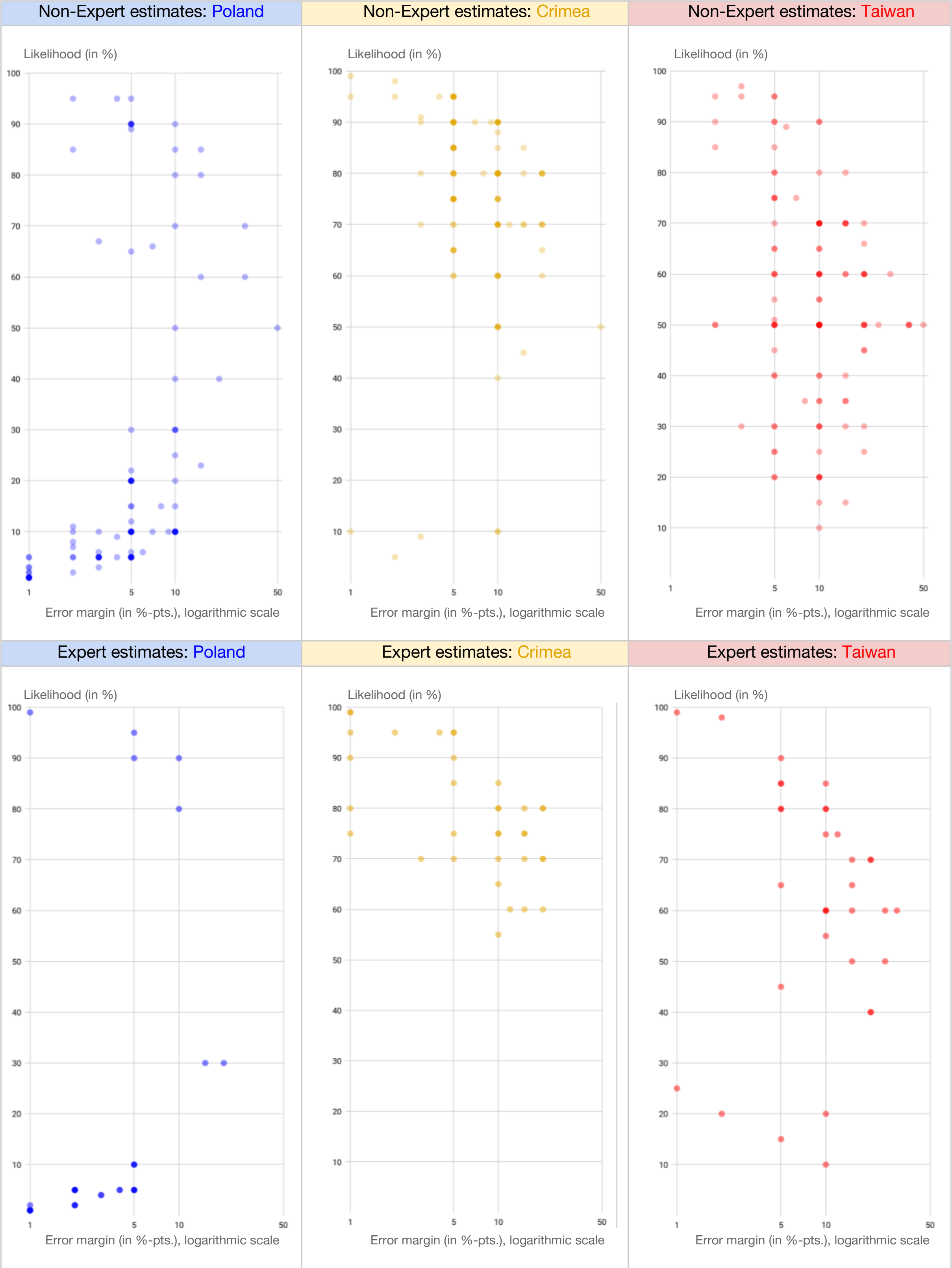
Table 4.4.1: *Two-sample (non-experts vs. experts) t-test statistics for participants' mean likelihood estimates*

Error margin	NATO (Poland)		Ukraine War (Crimea)		East Asian Sec. (Taiwan)	
t-test (ass. uneq. Var.)	Non-Experts	Experts	Non-Experts	Experts	Non-Experts	Experts
Mean	6.168067227	4.088235294	8.605042017	9.176470588	11.13445378	11.11764706
Variance	40.07320894	17.35561497	36.03760148	42.14973262	65.82922661	54.59180036
Observations	119	34	119	34	119	34
df	81		50		58	
t Stat	2.259607369		-0.460091741		0.01143863	
P (T<=t) one-tail	.013265712*		.323722485		.495456374	
P (T<=t) two-tail	.026531424*		.64744497		.990912748	

Table 4.4.2: *Two-sample (non-experts vs. experts) t-test statistics for participants' mean error margins*

Only exception: In the Poland/NATO scenario, experts' mean error margin (4.1 percentage points) is significantly smaller than that of the non-expert cohort (6.17 percentage points). That can be interpreted as the experts being more certain that Poland will likely not leave NATO in the next ten years than the non-experts.

[30] at the 0.05 significance level to reject the null hypothesis that both cohorts share the same true population mean



Non-Expert estimates: Poland

Non-Expert estimates: Crimea

Non-Expert estimates: Taiwan

Expert estimates: Poland

Expert estimates: Crimea

Expert estimates: Taiwan

Figure 4.7: *Non-experts' and experts' numeric estimates on likelihood (as a point estimate in percent) and analytic confidence (as an error margin in percentage point) for all three scenarios*

For the Poland/NATO scenario, numeric estimates are mostly below 50% likelihood (mean: 22.63) and error margins below 10 percentage points (mean: 5.7); with regards to the Crimea/Russia scenario, most estimates lie well above 50% likelihood (mean: 75.74) and most error margins between 5 and 10 percentage points (mean: 8.73); in the Taiwan/China scenario, likelihood estimates are spread across the whole range, roughly centred around the middle (mean: 56.21) and error margins are wider than in the other two assessments, mostly above 10 percentage points (mean: 11.13).

When comparing the VPE and numeric probability choices of each participant, some inconsistencies become apparent. Poland leaving NATO, for instance, was assessed by over 90% of non-experts as being “highly unlikely” or “unlikely”. However, there are suspiciously many numerical estimates above 50% likelihood; in some cases even going up to 95% (upper left graph of Fig. 4.7). Remarkably, the expert cohort also exhibits such outliers: No respondent in the expert cohort answered “likely” or “highly likely” in the first scenario (Poland/NATO) and yet, five participants in that expert cohort chose probabilities above 50% to express their assessment numerically (lower left graph of Fig. 4.7). By the second scenario, these inconsistent estimates disappear completely in the expert cohort and in the non-expert cohort they reduce (from 17 inconsistent estimates in the first scenario to 11 in the second scenario).

That is an important aspect that will be covered extensively later in the *secondary analytical aims*.

Based on the dataset of verbal probability assessments (verbal likelihood and level of analytic confidence) and the dataset of numeric estimates (point estimate on likelihood and margin of error for analytic confidence), the numeric interpretation of the VPEs can be computed (Fig. 4.8) – i.e. what odds did participants have in mind when assessing that something was "likely", for instance, and that they had "moderate confidence" in that assessment?

The non-expert likelihood interpretations look fairly consistent (upper left quadrant of Fig. 4.8): All four terms are ranked in the correct order on the scale from 0 to 100%, "unlikely" lies below 50% and "likely" above (the interquartile ranges of both terms even meet near the 50% mark). The interquartile ranges (IRQ) also don't overlap, indicating between-subject consistency in their numeric interpretations of the VPEs.

The expert likelihood interpretations, on the other hand, show some peculiarities (upper right quadrant of Fig. 4.8): While all four terms are also ordered correctly on the number line, both the median and the mean of the VPE "unlikely" lie above the 50% mark. Additionally, the IQRs of "likely" and "unlikely" overlap considerably, suggesting between-subject inconsistency in the interpretation of the two terms.

Since the majority of data points for the numeric interpretation of "unlikely" comes from the Poland/NATO scenario, these inconsistencies of the numeric interpretation of "unlikely" are probably a result of the outliers in the first scenario that were already discussed.[31] While these outliers are present in both cohorts, the expert cohort is the smaller dataset. Therefore, the outliers of the first scenario factor in more – shifting the overall median, quartiles and mean of "unlikely" up to the point of inconsistency.

[31] Almost all participants assessed the first scenario as "unlikely" or even "highly unlikely"; when expressing that assessment in numerical terms, however, some participants chose probabilities above 50%.

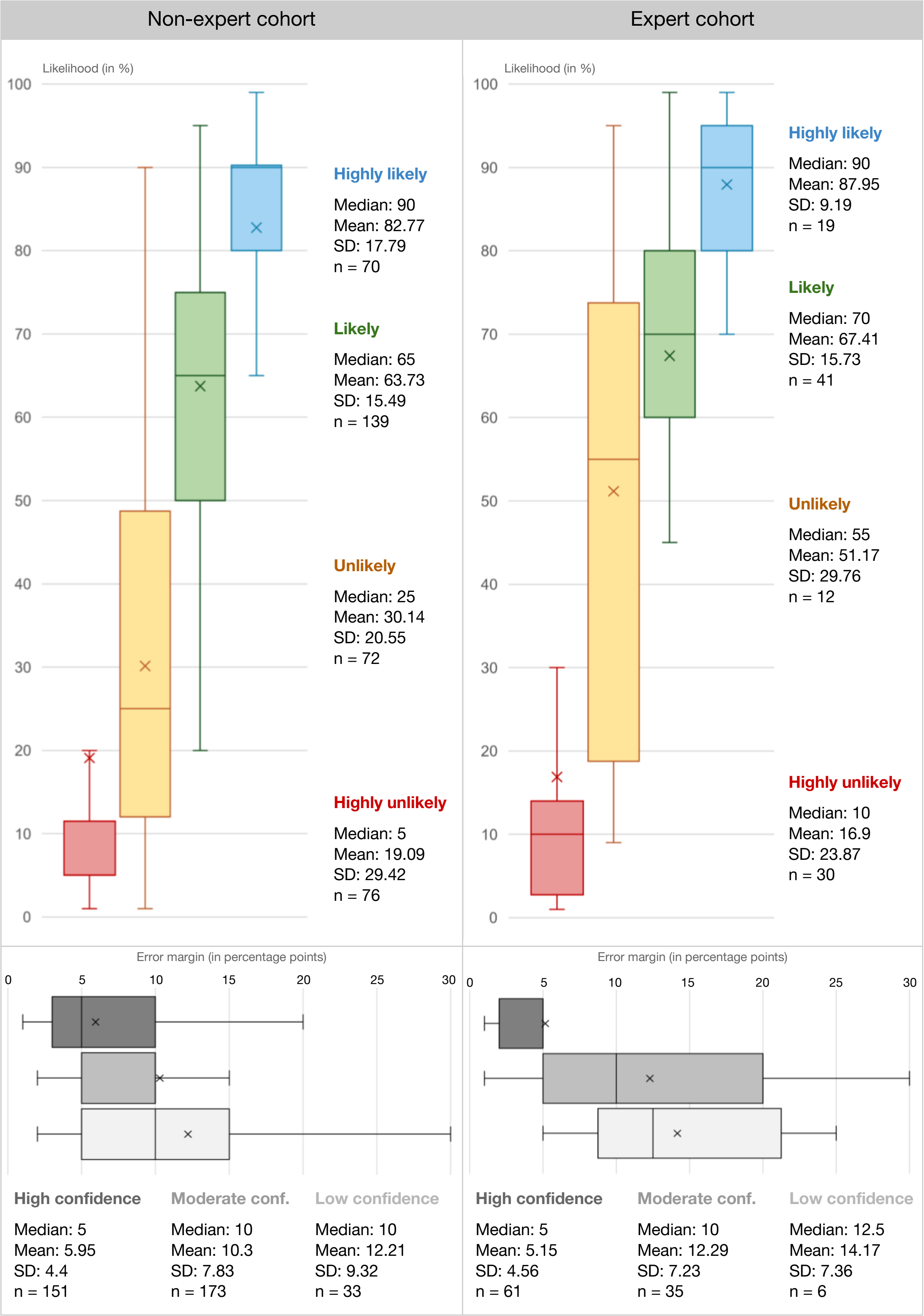


Figure 4.8: *Numeric interpretations of VPEs for likelihood (top) and analytic confidence (bottom)*

With respect to translating analytic confidence into a numeric error margin around the point estimate, both cohorts present some inconsistencies (lower part of Fig. 4.8).

For the non-experts (lower left quadrant of Fig 4.8), whether they had “low”, “moderate” or “high” analytic confidence in their assessment didn’t make much difference in the error margin they assigned. Most strikingly, the median estimate for “moderate” confidence (10 percentage points) is the same as for “low” confidence (also 10 percentage points).

While for the experts (lower right quadrant of Fig 4.8), at least the median estimates for “low” (12.5%-pts.), “moderate” (10%-pts.) and “high” (5%-pts.) are in a consistent order (the same goes for the mean estimates), they also have some calibration or consistency issues: The considerable overlap of the IRQs of “moderate” and “low” confidence, as well as the proximity of their means (12.29%-pts. / 14.17 %-pts.) and medians (10%-pts. / 12.5%-pts.) suggest either non-linear within-subject calibration or between-subject inconsistency.

Non-linear within-subject calibration means that these three confidence levels are not spaced equally. While “high” and “moderate” confidence are spaced 5 percentage points apart, it only takes 2.5 additional percentage points of error margin to make the jump from “moderate” to “low” confidence. Between-subject inconsistency means that multiple participants choose different verbal confidence levels (e.g. “moderate” or “low”) to express similar margins of error.

Due to the fact that considerably fewer expert participants used the confidence level “low” in their estimates (*n* = 6) compared to the usage of the term “moderate” (*n* = 35), one can assume that the latter, i.e. between-subject inconsistency, is more likely the case.

4.2 Analytic Statistics

Hypothesis 1 – Preference Divergence

> **H1:** When participants feel accountable for an estimate vis-à-vis a superior, they prefer verbal formats whereas they favour numerical formats when personal responsibility and hierarchical pressure is not present.

The first step to approach this hypothesis is to examine the preferences of the participants in the phase of the survey when they had to choose a format to submit their assessment in (phase 2) and compare that to the format they selected later in the experiment from a consumer's perspective (phase 4). The results for the three formats featured in the survey are visualised in Fig. 4.9.1 and the aggregated format preferences – verbal formats versus numerical – are shown in Fig. 4.9.2.

Both cohorts show some similarities in their preferences – especially, when considering the aggregated 'verbal-vs-numerical' categorisation of the data (Fig. 4.9.2). Not many expert participants chose the numeric range format in the production phase (only 15.69%). Instead, for half (50%) of expert assessments, the combination of VPE with an added verbal qualifier on analytic confidence was the format of choice. While this difference is not as stark with the non-expert cohort – the shares of preferences are more equally distributed among the three formats – it is still also present.

Comparing that to the consumer preferences, both cohorts show a quite remarkable change: The numeric range format becomes the most popular format among the three options (Fig. 4.9.1). In the aggregated categorisation (Fig. 4.9.2), both verbal and numerical formats now are head to head. So, already before performing any statistical analysis, the change in perspective clearly seems to have had some effect on format preference.

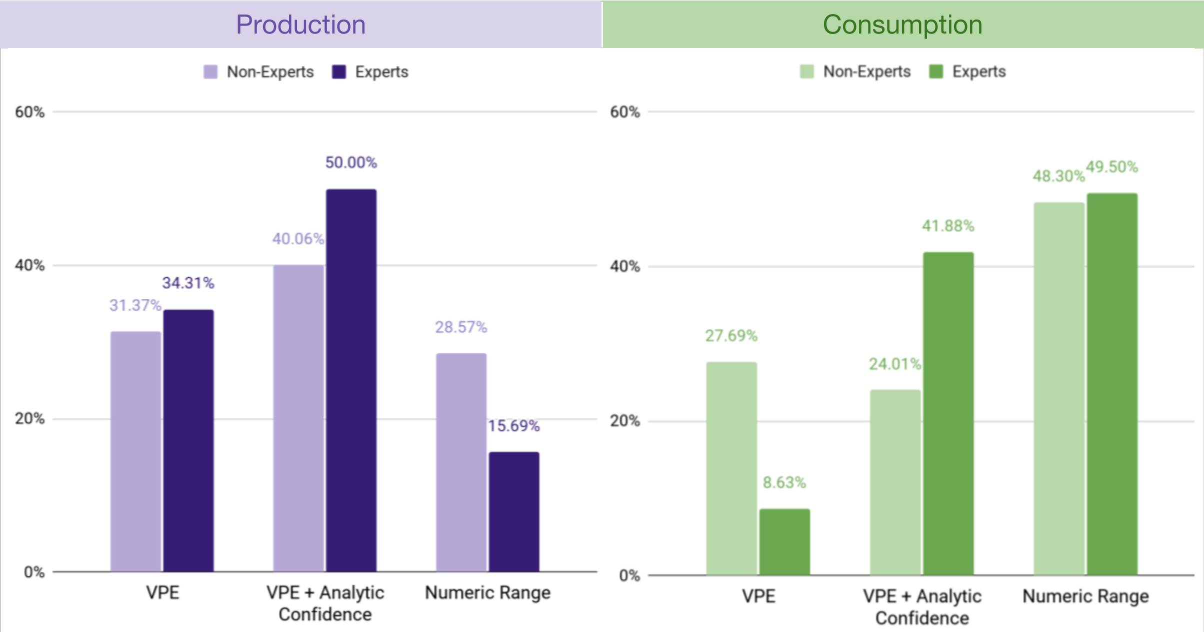


Figure 4.9.1: *Format preferences of non-experts and experts as estimate producers (left) and consumers (right), averaged across all three scenarios*

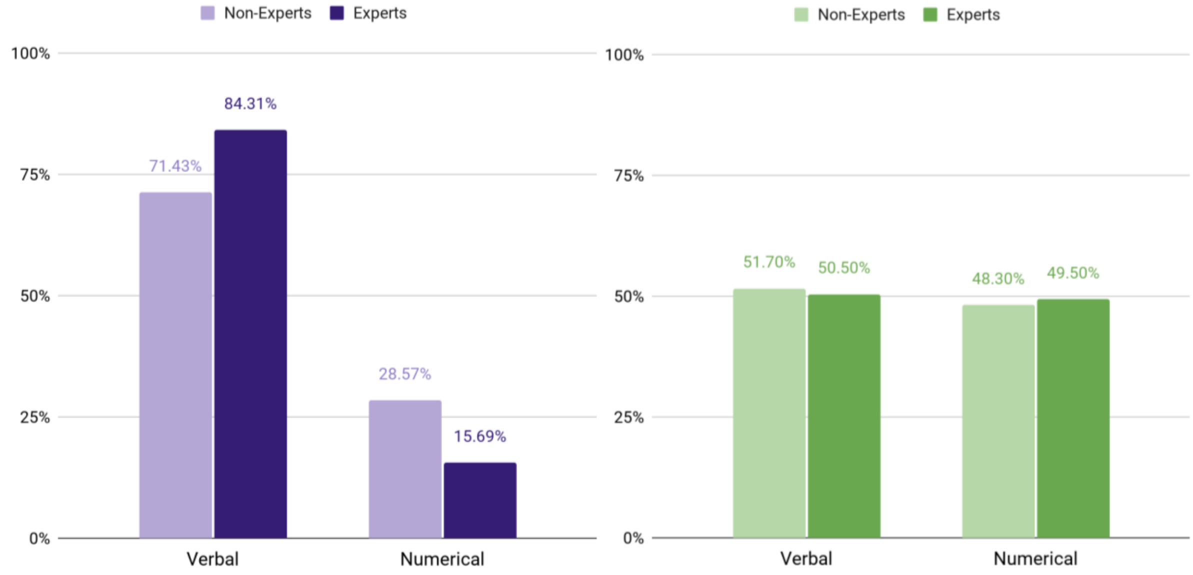


Figure 4.9.2: *Aggregated format preferences of non-experts and experts as estimate producers (left) and consumers (right), averaged across all three scenarios*

But before hypothesis 1 can be adequately tested, the data on the production preferences has to be manipulated. Due to the design of the survey, participants submitted assessments on *all three* subject areas (Poland/NATO, Crimea/Russia, Taiwan/China). However, they were asked about their format preference as a decision-maker only in the context of *one* of the three topics. To perform a true one-to-one comparison, for any given participant, only the producer preference must be considered in which the participant's consumer preference was elicited. For example, if a participant was asked about their consumer format preferences within the context of the Taiwan/China scenario, only the producer format preference for that very scenario will be considered. The resulting new shares in preferences (Fig. 4.10) don't seem to deviate that much from the previous preference distributions.

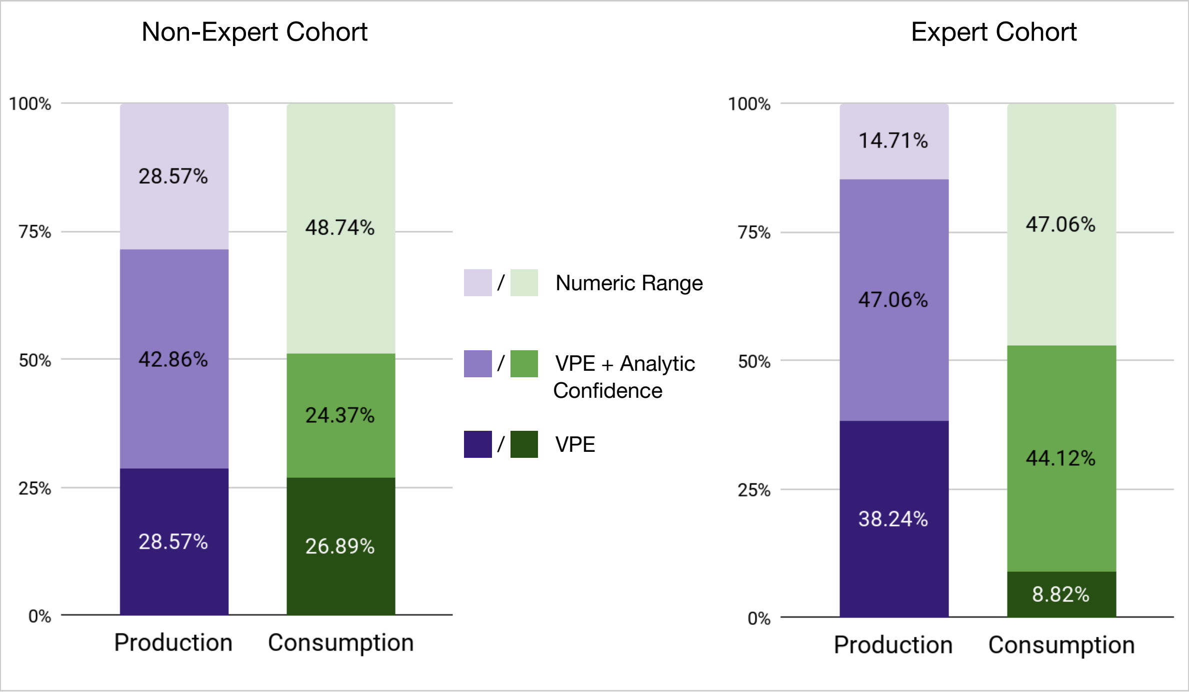


Figure 4.10: *Format preferences of non-experts (left) and experts (right), only considering the elicited production preferences in the scenario in which the consumer preference was also elicited in*

These new producer and consumer preferences – divided into verbal and numerical formats – can also be plotted in a contingency table, in which the columns indicate the producer preferences ($\text{Verbal}_{\text{Prod}}$ and $\text{Numerical}_{\text{Prod}}$) and the rows represent the consumer preferences ($\text{Verbal}_{\text{Prod}}$ and $\text{Numerical}_{\text{Prod}}$). This is done for all participants (Table 4.5.1) as well as for the two defined cohorts (Table 4.5.2 and 4.5.3).

N = 153	Verbal$_{\text{Prod}}$	Numerical$_{\text{Prod}}$
Verbal$_{\text{Cons}}$	65	14
Numerical$_{\text{Cons}}$	49	25

McNemar's test

$\chi^2 = 18.34920635$

df = 1

$p < .001$***

Effect size (phi coefficient)

$\Phi = 0.3463083211$

Table 4.5.1: *Contingency table for producer/consumer format preferences of all participants*

n = 119	Verbal$_{\text{Prod}}$	Numerical$_{\text{Prod}}$
Verbal$_{\text{Cons}}$	49	12
Numerical$_{\text{Cons}}$	36	22

McNemar's test

$\chi^2 = 11.02083333$

df = 1

$p < .001$***

Effect size (phi coefficient)

$\Phi = 0.3043222713$

Table 4.5.2: *Contingency table for producer/consumer format preferences of the non-expert cohort*

n = 34	Verbal$_{\text{Prod}}$	Numerical$_{\text{Prod}}$
Verbal$_{\text{Cons}}$	16	2
Numerical$_{\text{Cons}}$	12	3

McNemar's test

$\chi^2 = 5.785714286$

df = 1

$p = .01615693$*

Effect size (phi coefficient)

$\Phi = 0.4125143237$

Table 4.5.3: *Contingency table for producer/consumer format preferences of the expert cohort*

Based on these 2x2 contingency tables, McNemar's test can be performed – a test which is specifically designed for paired nominal data that was not obtained from independent samples (McNemar, 1947). That's precisely the case with this data, since preferences in the producer and consumer role were elicited from exactly the same participants. The test statistic for the corrected[32] McNemar's test is

[32] Edwards' (1948) continuity correction is used which is particularly important in dealing with the relatively small sample size of the expert cohort ($b + c < 25$).

approximated by the chi-squared distribution (with the frequencies in the four quadrants of the contingency table in reading order represented as a, b, c and d):

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c}$$

The null hypothesis of the McNemar test posits that the proportion of participants who prefer a verbal format as producers but the numerical format as consumers is equal to the proportion of participants whose preference changed the other way around (from numerical as producer to verbal as consumer).

$H_0: p_b = p_c$ (derived from $p_a + p_b = p_a + p_c$ and $p_c + p_d = p_b + p_d$)

Conversely, the alternative hypothesis states that preference changes in both directions are not proportionally equal – in this case: significantly more participants switched from verbal formats in the producer role to the numerical format in the consumer role.

Both the non-expert cohort (χ^2 = 11.02) as well as the overall sample (χ^2 = 18.35) show highly significant results (p < .001 in both cases). But also with the smaller cohort of experts (χ^2 = 5.79), the null hypothesis can still be rejected at a 0.05 significance level (p = .016). That means that the change in setting (from producer to consumer) resulted in a statistically significant change in format preference from verbal to numerical formats. According to Cohen (1988: 224ff.), a phi coefficient of $0.3 \leq \phi < 0.5$ constitutes a medium effect size. All three cases – the overall dataset (ϕ = 0.3463), the non-expert cohort (ϕ = 0.3043) and the expert cohort (ϕ = 0.4125) – fall into that effect size category. However, that is not enough for the numerical format to surpass the verbal formats in the consumer setting – instead, consumer preferences are pretty evenly split between these two format types (Fig. 4.10).

The survey also asked participants about the reasons behind their choice of format. Half (50%) of the respondents who chose a verbal format selected “It leaves room for ambiguity / It is not overly-specific” as their rationale. Conversely, the largest share (44.94%) of participants that opted for the numerical format reported precision as the quality guiding their choice. Table 4.6 plots the absolute frequencies of those respondents who reported either quality to have had an influence on their format preference.

n = 102	Ambiguity	Precision
Verbal	32	15
Numerical	15	40

Pearson’s chi-squared

χ^2 = 16.98963294

df = 2

p < .001***

Table 4.6: *Contingency table plotting absolute frequencies of consumer format and ambiguity/precision preferences*

Pearson’s chi-squared test reveals that these qualities and the format preference are not independent – on the contrary: with *p* < .001 the association is highly significant. The numerical format is chosen for its precision, while the verbal format is selected by respondents who value the opposite; to avoid misplaced overprecision and/or because they value its deliberate ambiguity. These results are also significant at a 0.01 significance level in the expert (*p* = .00943400379) and the non-expert cohort (*p* < .001). Two other qualities that were offered as answer options, comprehensibility and transparency (about the assessment’s uncertainties), did not show significant associations with format preference. Among respondents preferring a verbal format, 40.2% cited comprehensibility as a guiding quality and 29.89% comprehensibility. Among participants with a preference for the numeric format, 27.5% selected comprehensibility and 35% transparency. Note that the choice of quality was not exclusive; participants could select multiple qualities as influential for their format preference.

Hypothesis 2 – Congruence Principle

H2: Estimate producers are more likely to choose numerical formats the more certain the assessment is.

Based on the data collected through the survey, there are several possible options to measure the underlying certainty of a participant's assessment. The most straightforward method would be to just use the analytic confidence levels that participants provided with each assessment as a measure of their certainty. As a second – and related – option, the error margin, as the numeric translation of a participant's analytic confidence, could be used.

But also the likelihood assessment can function as a basis to measure an estimate's underlying certainty. Assessments deeming something either 0% likely or 100% likely are maximally certain – suggesting either impossibility or perfect certainty. In that logic, a 50% likelihood assessment, i.e. 50/50 odds, would be maximally uncertain. Based on that conceptualisation, the 'distance to the extremes' (i.e. how far an assessment is from either 0% or 100%) is a further possible measure of uncertainty. Something similar can be done for the verbal probability assessments, assuming that "highly unlikely" and "highly likely" show more certainty than "unlikely" and "likely".

Lastly, the three different scenarios could be used as categories to delineate different levels of certainty. As the submitted assessments discussed previously show, each scenario yielded differing results in terms of likelihood and analytic confidence.[33]

[33] See '*Submitted Assessments*' in *4.1 Descriptive Statistics*

In summary, these are the five possible measures of uncertainty that are tested for their association with participants' format preferences:

1. Analytic confidence, verbal
2. Analytic confidence, numerical (error margin)
3. 'Distance to the extremes', numerical
4. 'Distance to the 'extremes', verbal
5. Scenario

To test whether the verbal level of analytic confidence is associated with the format type chosen to submit the assessments (verbal or numerical), contingency tables are used again. Table 4.7.1 plots the frequencies of format choice and confidence level in the whole dataset; the other two tables do the same, but only for the non-expert (Table 4.7.2) and expert cohort (Table 4.7.3).[34]

N = 153	High conf.	Moderate conf.	Low conf.	*CMH test* M^2 = 15.247
Verbal	175	139	27	df = 2
Numerical	37	69	12	$p < .001$***

Table 4.7.1: *Aggregated contingency table plotting format choices dependent on analytic confidence level of all participants*

n = 119	High conf.	Moderate conf.	Low conf.	*CMH test* M^2 = 13.207
Verbal	122	111	22	df = 2
Numerical	29	62	11	$p = .001355$**

Table 4.7.2: *Aggregated contingency table plotting format choices dependent on analytic confidence level of the non-expert cohort*

[34] The table sums are always three times the size of the number of participants since each respondent submitted three estimative judgements.

				CMH test
n = 34	High conf.	Moderate conf.	Low conf.	M^2 = 2.2857
Verbal	53	28	5	df = 2
Numerical	8	7	1	*p* = .3189

Table 4.6.3: *Aggregated contingency table plotting format choices dependent on analytic confidence level of the expert cohort*

To assess whether there is a statistically significant association between the level of analytic confidence and the choice of format, Pearson's chi-squared test could be used. That, however, would disregard the fact that the data was obtained using repeated measures – each participant submitted three assessments. Therefore, the assumption of independence does not hold and there might be within-subject correlations, which an ordinary chi-squared test wouldn't account for. To control for a possible 'participant effect', the generalised[35] Cochran-Mantel-Haenszel (CMH) test is used: It tests the association between two categorical variables – similar to other chi-squared tests – but accounts for a third confounding categorical variable.

The CMH test formulates the null hypothesis through odds ratios. If, for each participant (i), the level of analytic confidence (X) and the choice of format (Y) are conditionally independent, then the odds ratios (θ) should all be equal to 1 (Agresti, 2007: 29). Therefore:

$$H_0: \theta_{XY(i=1)} = \theta_{XY(i=2)} = \ldots = \theta_{XY(i=153)} = 1$$

Conversely, the alternative hypothesis states that confidence level and format choice are not conditionally independent – i.e. the odds ratios for each participant are not all equal to 1. The generalised CMH test outputs a test statistic M^2 which is close to

[35] The initial version of the test was designed for 2x2 contingency tables (Cochran, 1954; Mantel/Haenszel, 1959). Since, in this case, we deal with 3x2 contingency tables, the generalised CMH test is employed, which can handle larger contingency tables (Somes, 1986).

zero when all $\theta_{XY(k)} = 1$ and gets larger if some or all $\theta_{XY(k)} > 1$ or $\theta_{XY(k)} < 1$, asymptotically following the chi-squared distribution with, in this case, 2 degrees of freedom (Somes, 1986).

To perform this test, the data is structured into 3x2 contingency tables (plotting confidence level and format preference) for each of the 153 participants, resulting in a three-dimensional 3x2x153 table. The same is done for the two cohorts which produces a 3x2x34 table for the experts and a 3x2x119 table for the non-experts. The resulting test statistics M^2 and the corresponding p-values are provided in Tables 4.7.1 through 4.7.3.

Running the CMH test on the whole dataset as well as on the non-expert cohort produces statistically highly significant results ($p < .01$ in both cases). Thus, the null hypothesis that confidence level and format preference are independent can be wholeheartedly rejected in those two cases. Interestingly, the test statistic for the expert cohort comes out quite small ($M^2 = 2.29$), missing the 0.05 threshold for statistical significance by a lot ($p = .3189$).

Cramér's *V* is used to evaluate the effect size in the two statistically significant cases (overall dataset and non-expert cohort). Cramér's *V* can take values between 0 (no association of the two variables) and 1 (perfect association). The effect sizes are fairly low, both in the overall dataset (Cramér's $V = 0.176$) as well as the non-expert cohort (Cramér's $V = 0.178$).[36] It is important to note that Cramér's *V* does not account for repeated measures (similar to Pearson's chi-squared test it assumes independence between the variables' frequencies), which can potentially inflate the association values. But especially since these potentially inflated values for Cramér's

[36] Running Cramér's *V* on the expert cohort's dataset produced – as expected – an even lower effect size (Cramér's $V = 0.089$).

V are this small, that strongly suggests that the association between confidence level and format preference – though significant – is weak.

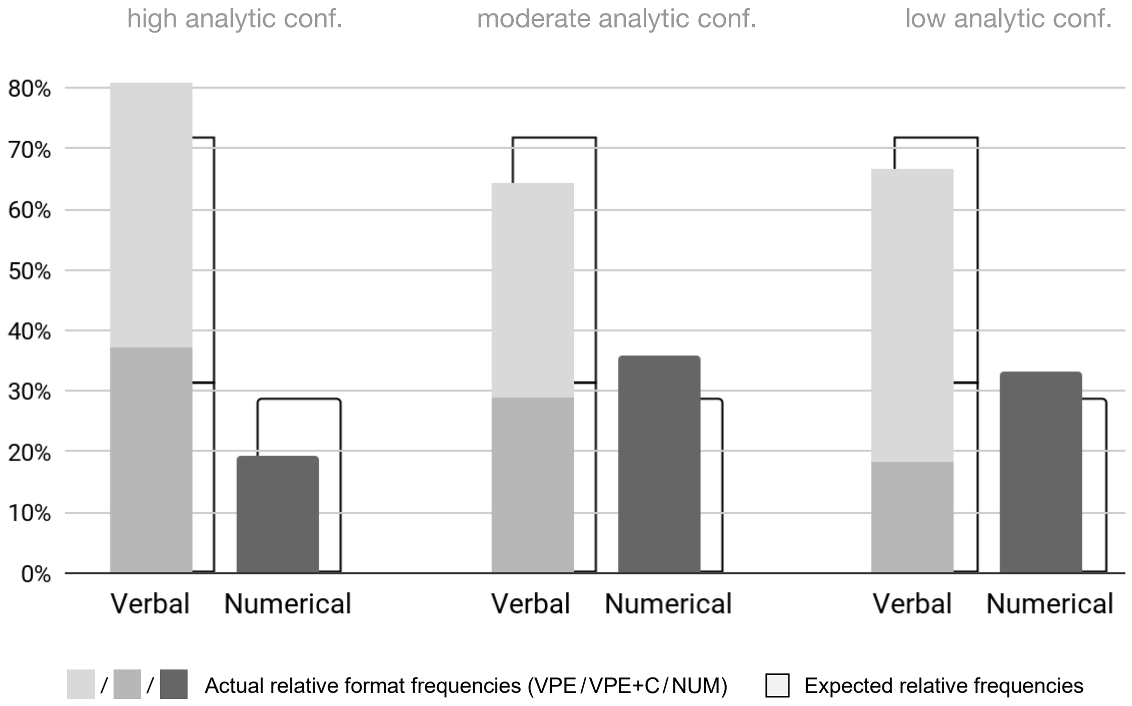


Figure 4.11.1: *Format choices for each level of analytic confidence of the* <u>*non-expert cohort*</u>

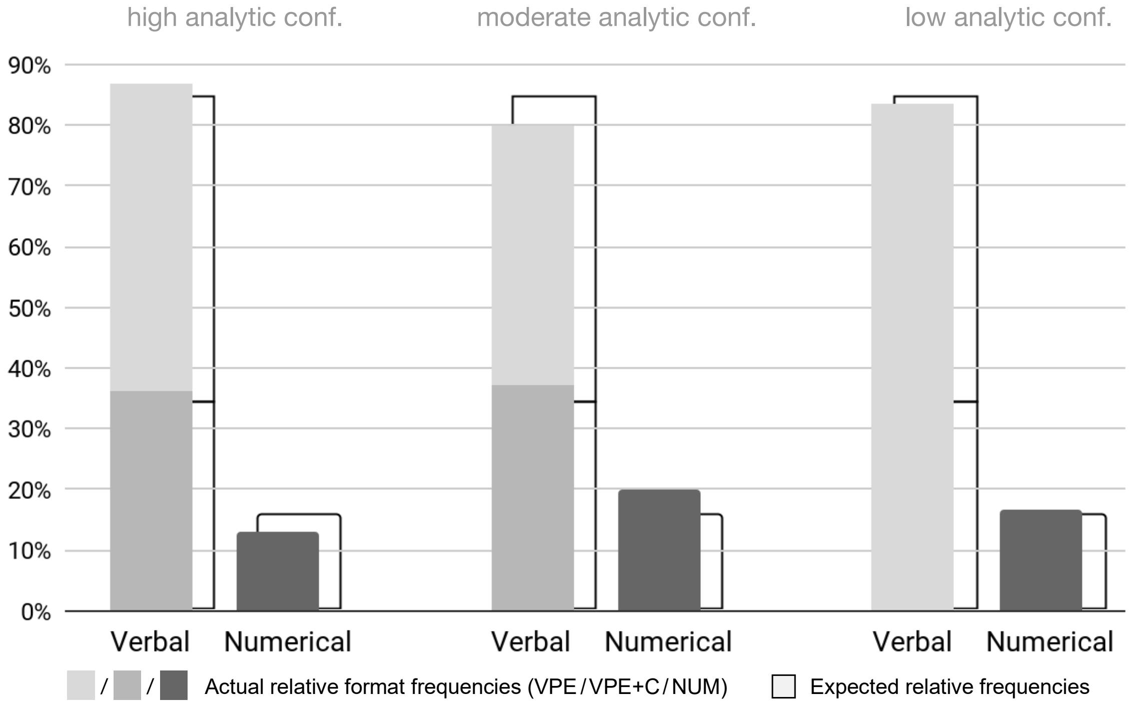


Figure 4.11.2: *Format choices for each level of analytic confidence of the* <u>*expert cohort*</u>

To assess the nature of the association, the relative frequencies of both format types (verbal and numerical) in all three confidence levels, are plotted against their expected[37] relative frequencies both for the non-expert cohort (Fig. 4.11.1) and the expert cohort (Fig. 4.11.2). The most striking deviations from the expected relative frequencies can be seen when comparing "high" and "moderate" confidence assessments: the numerical format underperforms in the former while performing better than expected in the latter assessment category. The overperformance of the numerical format continues in the "low analytic confidence" group of assessments. What also becomes apparent from comparing both cohorts, however, is why the non-expert cohort produced a statistically significant association while the expert-cohort's dataset did not. In the expert cohort, the deviations from the expected relative frequencies are rather minor; in the non-expert cohort, these deviations are more pronounced. Nevertheless, it also makes sense that the association measure (Cramér's V) returned small effect sizes, since – even in the non-expert cohort – the deviations from the expected relative frequencies don't shift the overall distribution significantly.

Since the error margin is supposed to basically be the numerical translation of the analytic confidence level, it could be expected that it produces similar results as the verbal confidence levels. In order to assess the association of the error margin on format choice, while controlling for repeated measures, a mixed effects logistic regression model is used. The error margin that respondents provided with their assessments is the fixed effect in the model. Each participant is numbered and that *Participant ID* (running from 1 to 153) is treated as a random effect to account for the repeated measures (each participant contributed three assessments to the dataset). The dependent variable is the format preference coded as a binary variable (0 for

[37] under the assumption of independence

verbal and 1 for numerical). This analysis is conducted on the whole dataset (N = 459 observations from 153 participants) as well as on the two cohorts – the experts (n = 102 observations from 34 participants) and the non-experts (n = 357 observations from 119 participants).

Generalised linear mixed model fit by maximum likelihood (Laplace Approximation)

Overall dataset				
Predictor	**Estimate (β)**	**Std. Error**	***z*-value**	***p*-value**
(Intercept)	-1.784225	0.338186	-5.276	< .001***
Error Margin	0.009222	0.022936	0.402	.688
Random effects	**Variance**	**Std. Dev.**		
Participant ID	3.353	1.831		

Table 4.8.1:
Statistics for the mixed-effects logistic regression model, all participants

Non-expert cohort				
Predictor	**Estimate (β)**	**Std. Error**	***z*-value**	***p*-value**
(Intercept)	-1.435913	0.365887	-3.924	< .001***
Error Margin	-0.007425	0.026358	-0.282	.778
Random effects	**Variance**	**Std. Dev.**		
Participant ID	3.491	1.868		

Table 4.8.2:
Statistics for the mixed-effects logistic regression model, non-expert cohort

Expert cohort				
Predictor	**Estimate (β)**	**Std. Error**	***z*-value**	***p*-value**
(Intercept)	-2.77669	0.80111	-3.466	< .001***
Error Margin	0.06205	0.04827	1.285	.1987
Random effects	**Variance**	**Std. Dev.**		
Participant ID	1.799	1.341		

Table 4.8.3:
Statistics for the mixed-effects logistic regression model, expert cohort

None of the three models show statistical significance for the error margin as a predictor of format choice. Both the overall dataset (Table 4.8.1) and the non-expert cohort (Table 4.8.2) miss the 0.05 significance threshold by a lot. With the expert cohort (Table 4.8.3), the *p*-value of the error margin as predictor – while being considerably lower than with the other cohort – is still way above the required $p \leq 0.05$ to reject the null hypothesis[38].

The third possible measure of uncertainty being tested is the likelihoods' distance to the extremes, 0% or 100%. It is calculated as follows (where x is the submitted likelihood estimate and D the distance to the nearest extreme):

$$D_i = 50 - |x_i - 50|$$

Similar to the mean error margins, a mixed effects logistic regression model is used to assess the effect of the independent variable *D* (distance to the nearest extreme) on the dependent variable *format preference* (again, coded as 0 for verbal and 1 for numerical), while treating *Participant ID* as a random effect. And, as with the error margins, the results fall short of statistical significance both with the overall dataset (Table 4.9.1) and with the two cohorts (Table 4.9.2 and 4.9.3).

Generalised linear mixed model fit by maximum likelihood (Laplace Approximation)

Overall dataset				
Predictor	**Estimate (β)**	**Std. Error**	***z*-value**	***p*-value**
(Intercept)	-1.89298	0.33767	-5.606	< .001***
Dist. to the extremes (D)	0.01391	0.01432	0.971	.331
Random effects	**Variance**	**Std. Dev.**		
Participant ID	3.446	1.856		

Table 4.9.1:
Statistics for the mixed-effects logistic regression model, all participants

[38] That the predictor *error margin* has no effect on the dependent variable *format preference*

Non-Experts				
Predictor	**Estimate (β)**	**Std. Error**	***z*-value**	***p*-value**
(Intercept)	-1.71541	0.36734	-4.670	< .001***
Dist. to the extremes (D)	0.01913	0.01963	0.975	.33
Random effects	**Variance**	**Std. Dev.**		
Participant ID	3.568	1.889		

Table 4.9.2:
Statistics for the mixed-effects logistic regression model, non-expert cohort

Experts				
Predictor	**Estimate (β)**	**Std. Error**	***z*-value**	***p*-value**
(Intercept)	-2.71971	0.80805	-3.366	< .001***
Dist. to the extremes (D)	0.02416	0.02247	1.075	.282181
Random effects	**Variance**	**Std. Dev.**		
Participant ID	1.837	1.355		

Table 4.9.3:
Statistics for the mixed-effects logistic regression model, expert cohort

For the verbal version of the ‘distance to the extremes’-measure of uncertainty (assuming “highly unlikely” and “highly likely” suggest more certainty than “unlikely” and “likely”), the Mantel Haenszel’s statistic[39] is used. But also here, no statistically significant results were produced (Table 4.10.1 through 4.10.3).

N = 459	Highly un-/likely	Un-/likely
Verbal	147	194
Numerical	48	70

Mantel-Haenszel statistic

$\chi^2 = 0.30867$

df = 1

p = .5785

Table 4.10.1: *Contingency table plotting format choices dependent on verbal ‘distance to the extremes’ of all participants*

[39] The Mantel-Haenszel statistic is a variant of the generalised CMH test that was used earlier in the analysis. It is suited for 2x2 contingency tables while controlling for a third confounding variable and is approximated by the chi-squared distribution (Somes, 1986; Agresti, 2007: 114f.).

n = 357	Highly un-/likely	Un-/likely
Verbal	105	150
Numerical	41	61

Mantel-Haenszel statistic

χ^2 = 0.082237

df = 1

p = .7743

Table 4.10.2: *Contingency table plotting format choices dependent on verbal 'distance to the extremes' of the non-expert cohort*

n = 102	Highly un-/likely	Un-/likely
Verbal	42	44
Numerical	7	9

Mantel-Haenszel statistic

χ^2 = 0.10227

df = 1

p = .7491

Table 4.10.3: *Contingency table plotting format choices dependent on verbal 'distance to the extremes' of the expert cohort*

Lastly, it was tested[40] whether the scenario – and possibly perceived differences in the certainty of these scenarios – had an effect on format preferences. Neither in the overall sample (Table 4.11.1), nor in the two cohorts (Table 4.11.2 and 4.11.3) did the scenario show a statistically significant effect on the choice of format.

N = 459	Poland	Crimea	Taiwan
Verbal	115	117	109
Numerical	38	36	44

CMH test

M^2 = 1.8571

df = 2

p = .3951

Table 4.11.1: *Aggregated contingency table plotting format choices of all participants dependent on the scenario*

n = 357	Poland	Crimea	Taiwan
Verbal	85	86	84
Numerical	34	33	35

CMH test

M^2 = 0.13333

df = 2

p = .9355

Table 4.11.2: *Aggregated contingency table plotting format choices of the non-expert cohort dependent on the scenario*

[40] Again using the CMH test for 3x2 contingency tables while controlling for repeated measures

n = 102	Poland	Crimea	Taiwan
Verbal	30	31	25
Numerical	4	3	9

CMH test

M^2 = 5.6364

df = 2

p = .05971

Table 4.11.3: *Aggregated contingency table plotting format choices of the expert cohort dependent on the scenario*

It has to be noted, however, that the expert cohort comes quite close to a statistically significant association between the scenario and the choice of format (*p* = .05971). Why that is, becomes quite apparent when plotting the format preferences per scenario (Fig. 4.12): In the expert cohort, the numerical format sees a considerable uptick in the Taiwan/China scenario, relative to its unpopularity in the other two scenarios. The non-experts' format preferences, on the other hand, stay pretty stable across scenarios.

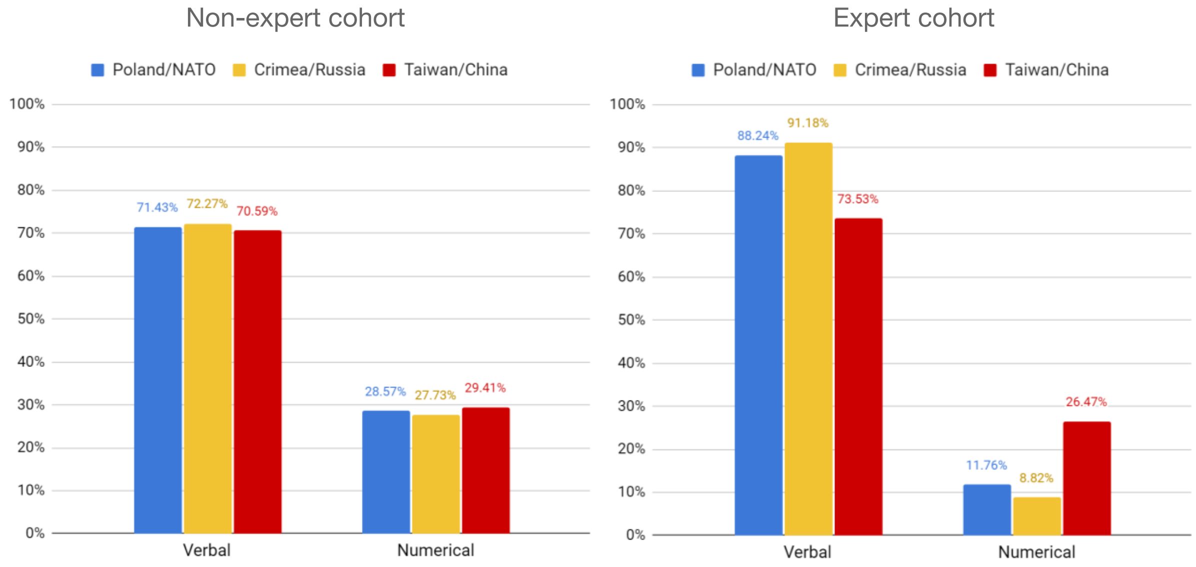


Figure 4.12: *Non-experts' (left) and experts' (right) producer format preferences across scenarios*

In summary, estimate producers are not more likely to choose numeric formats the more certain an assessment is – in fact, the opposite seems to be the case: Preference for the numeric range format was significantly higher for assessments with "moderate" or "low" analytic confidence than for "high" confidence. That directly contradicts the congruence principle.

Hypothesis 3 – Illusion of Rigour

H3: Consumers of estimative judgements associate numerical formats with a higher expertise of the producer behind the assessment.

The data to test that hypothesis was tested in phase 3 of the survey. Out of the 153 respondents, all but one completed that part of the experiment.[41] Therefore, the expert cohort is reduced by one, compared to the other parts of the survey.

The overall results from phase 3 are plotted in Table 4.12.1. The data for the five format options is then aggregated in Table 4.12.2, distinguishing between verbal and numerical formats. Columns represent the three different scenarios as well as the aggregated result across all three scenarios; the different format options that participants could choose from are listed in the rows. The value in the left upper corner of each cell is the absolute frequency that a format was chosen in a given scenario. The grey value in the upper centre of each cell is the expected frequency under the null hypothesis that all options are equally likely to be picked. The value in the upper right corner is the delta between the expected and the observed frequency (red if negative, green if positive). The *p*-value in the lower half of each cell is the probability of finding a frequency equal to or smaller than the observed frequency using the binomial distribution. For cases where the observed value is greater than the expected value, the inverse probability of the binomial distribution is used. Cells are shaded green (for values greater than the expected value) or red (for values smaller than the expected value) if the *p*-value falls below the 0.05 significance threshold to reject the null hypothesis.

[41] The participant's self-reported level of expertise was the highest, "recognised expert knowledge", for all three subject areas. Therefore, this respondent was not shown phase 3 because the conditions that the experiment would have demanded couldn't be met.

All (*N* = 304)	Poland/NATO	Crimea/Russia	Taiwan/China	Total
VPE$_{original}$	**24** 22.8 +1.2 *p* = .4259756278	**20** 21.2 -1.2 *p* = .4419614709	**16** 16.8 -0.8 *p* = .4783243701	**60** 60.8 -0.8 *p* = .4885722879
VPE$_{altered}$	**20** 22.8 -2.8 *p* = .3010532777	**15** 21.2 -6.2 *p* = .07912483989	**9** 16.8 -7.8 *p* = .01798137509*	**44** 60.8 -16.8 *p* = .007939623421**
VPE+C	**31** 22.8 +8.2 *p* = .03931218265*	**23** 21.2 +1.8 *p* = .3679366747	**22** 16.8 +5.2 *p* = .1024273134	**76** 60.8 +15.2 *p* =.01950736341*
NUM$_{original}$	**16** 22.8 -6.8 *p* = .06574088561	**18** 21.2 -3.2 *p* = .2606601469	**21** 16.8 +4.2 *p* = .1562669632	**55** 60.8 -5.8 *p* = .2254758363
NUM$_{altered}$	**23** 22.8 +0.2 *p* = .5186264007	**30** 21.2 +8.8 *p* = .02534552058*	**16** 16.8 -0.8 *p* = .4783243701	**69** 60.8 +8.2 *p* = .1353528743

Table 4.12.1: *Distribution of 'expert assessments', all participants*

All (*N* = 304)	Poland/NATO	Crimea/Russia	Taiwan/China	Total
Verbal	**75** 68.4 +6.6 *p* = .1212194908	**58** 63.6 -5.6 *p* = .1559891202	**47** 50.4 -3.4 *p* = .2578891295	**180** 182.4 -2.4 *p* = .4105643917
Numerical	**39** 45.6 -6.6 *p* = .1212194908	**48** 42.4 +5.6 *p* = .1559891202	**37** 33.6 +3.4 *p* = .2578891295	**124** 121.6 +2.4 *p* = .4105643917

Table 4.12.2: *Aggregated distribution of 'expert assessments', all participants*

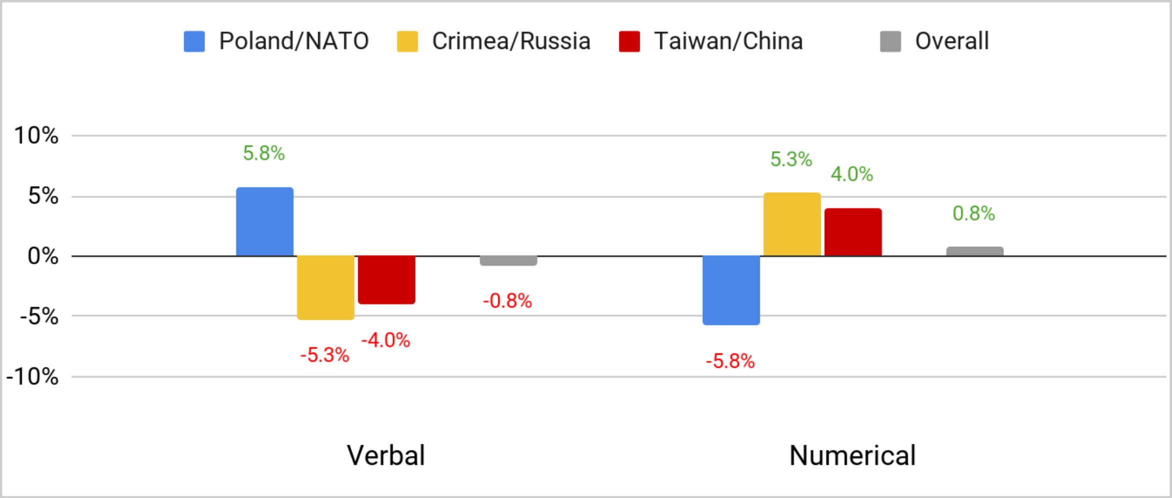


Figure 4.13: *Over- or underperformance of the verbal and numerical formats across scenarios (relative to the expected frequencies), all participants*

The experiment's results do not show a clear trend that numerical formats are significantly more likely to be associated with expertise (Table 4.12.2). In the aggregated results (verbal vs. numerical), over- or underperformance varies across scenarios, yielding an overall result close to the expected values (Fig. 4.13).

Across all three scenarios and all five format options, the altered VPE format underperformed a bit while the combination of VPE and analytic confidence level was chosen more often than stochastically expected (Table 4.12.1).

Breaking these results down reveals some differences between the non-expert and expert cohort. For the non-experts, the VPE+C format was chosen significantly more often to be the expert assessment (Table 4.13.1), possibly accounting for this format's overperformance in the whole dataset. In the aggregation, however, neither verbal nor numerical formats are chosen significantly more or less often (Table 4.13.2). Instead, their performance shows great volatility across scenarios (Fig. 4.14).

Non-Experts	Poland/NATO	Crimea/Russia	Taiwan/China	Total
VPE$_{original}$	**21** 20.4 +0.6 *p* = .4802094639	**15** 17.2 -2.2 *p* = .3313419575	**11** 10 +1 *p* = .4164405815	**47** 47.6 -0.6 *p* = .5000217126
VPE$_{altered}$	**20** 20.4 -0.4 *p* = .5197905361	**12** 17.2 -5.2 *p* = .09899782408	**7** 10 -3 *p* = .1904098116	**39** 47.6 -8.6 *p* = .09243380889
VPE+C	**28** 20.4 +7.6 *p* = .04322252774*	**17** 17.2 -0.2 *p* = .5428982334	**16** 10 +6 *p* = .03080342278*	**61** 47.6 +13.4 *p* = .02053826042*
NUM$_{original}$	**13** 20.4 -7.4 *p* = .03863202787*	**17** 17.2 -0.2 *p* = .5428982334	**9** 10 -1 *p* = .4437404133	**39** 47.6 -8.6 *p* = .09243380889
NUM$_{altered}$	**20** 20.4 -0.4 *p* = .5197905361	**25** 17.2 +7.8 *p* = .02838316451*	**7** 10 -3 *p* = .1904098116	**52** 47.6 +4.4 *p* = .26055298159

Table 4.13.1: *Distribution of 'expert assessments', non-expert cohort*

Non-Experts	Poland/NATO	Crimea/Russia	Taiwan/China	Total
Verbal	**69** 61.2 +7.8 *p* = .06874883444	**44** 51.6 -7.6 *p* = .06001374698	**34** 30 +4 *p* = .1560906047	**147** 142.8 +4.2 *p* = .3134609831
Numerical	**33** 40.8 -7.8 *p* = .06874883444	**42** 34.4 +7.6 *p* = .06001374698	**16** 20 -4 *p* = .1560906047	**91** 95.2 -4.2 *p* = .3134609831

Table 4.13.2: *Aggregated distribution of 'expert assessments', non-expert cohort*

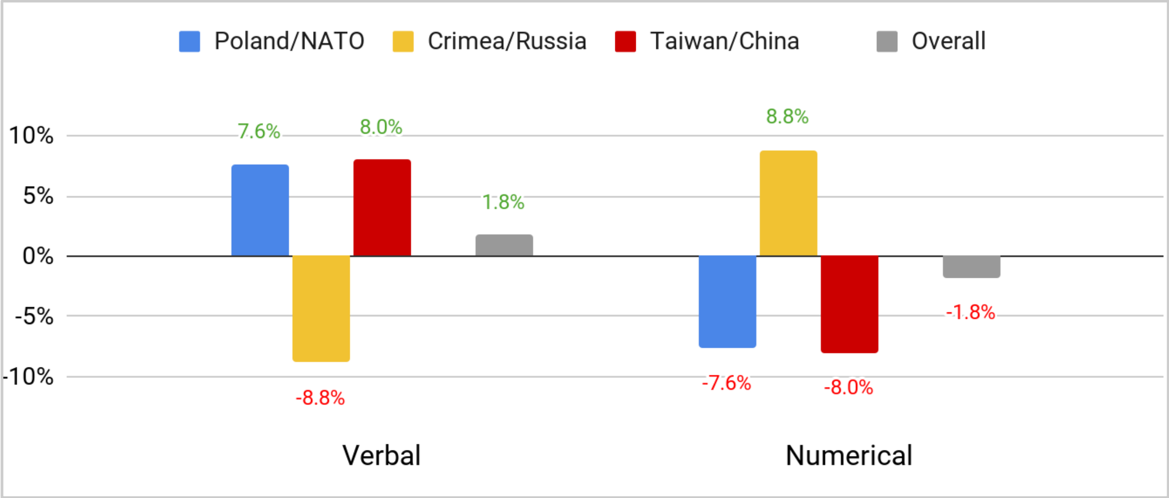


Figure 4.14: *Over- or underperformance of the verbal and numerical formats across scenarios (relative to the expected frequencies), non-expert cohort*

The expert cohort's results are particularly interesting because they come perhaps the closest to supporting the illusion of rigour hypothesis.

First, the altered VPE format underperforms in the expert cohort. In the context of the Taiwan/China scenario, this underperformance's *p*-value falls below the 0.05 threshold for statistical significance (Table 4.14.1). But also in the Poland/NATO scenario, $\text{VPE}_{\text{altered}}$ performs worse than statistically expected, not once being chosen as a supposed 'expert assessment' and almost reaching statistical significance (*p* = .069). This format's bad performance with the expert cohort might likely be the reason for its underperformance in the overall dataset.

Secondly, the expert cohort seems to have a certain bias towards the numerical format when it comes to perceived expertise. This overperformance of the numerical formats and underperformance of the verbal ones with the expert cohort particularly shows in the aggregated view of the data (Table 4.14.2 and Fig. 4.15).

Experts	Poland/NATO	Crimea/Russia	Taiwan/China	Total
$\text{VPE}_{\text{original}}$	**3** 2.4 +0.6 *p* = .4416542515	**5** 4 +1 *p* = .3703517361	**5** 6.8 -1.8 *p* = .2996488334	**13** 13.2 -0.2 *p* = .5489258102
$\text{VPE}_{\text{altered}}$	**0** 2.4 -2.4 *p* = .06871947674	**3** 4 -1 *p* = .411448862	**2** 6.8 -4.8 *p* = .02259587195*	**5** 13.2 -8.2 *p* = .004985261644**
VPE+C	**3** 2.4 +0.6 *p* = .4416542515	**6** 4 +2 *p* = .1957922145	**6** 6.8 -0.8 *p* = .4661398845	**15** 13.2 +1.8 *p* = .3352186758
$\text{NUM}_{\text{original}}$	**3** 2.4 +0.6 *p* = .4416542515	**1** 4 -3 *p* = .06917529028	**12** 6.8 +5.2 *p* = .02743998132*	**16** 13.2 +2.8 *p* = .2348105638
$\text{NUM}_{\text{altered}}$	**3** 2.4 +0.6 *p* = .4416542515	**5** 4 +1 *p* = .3703517361	**9** 6.8 +2.2 *p* = .2268922401	**17** 13.2 +3.8 *p* = .1547978494

Table 4.14.1: *Distribution of 'expert assessments', expert cohort*

Experts	Poland/NATO	Crimea/Russia	Taiwan/China	Total
Verbal	**6** 7.2 -1.2 *p* = .3347914424	**14** 12 +2 *p* = .2500106719	**13** 20.4 -7.4 *p* = .008460422608**	**33** 39.6 -6.6 *p* = .06371880592
Numerical	**6** 4.8 +1.2 *p* = .3347914424	**6** 8 -2 *p* = .2500106719	**21** 13.6 +7.4 *p* = .008460422608**	**33** 26.4 +6.6 *p* = .06371880592

Table 4.14.2: *Aggregated distribution of 'expert assessments', expert cohort*

However, it is a bit peculiar that these significant deviations from the expected frequencies only occur in the Taiwan/China context – the scenario in which the algorithm to generate the altered formats differed from the other two scenarios. Thus, the underperformance of the altered VPE format and the overperformance of the unaltered numeric format might be artefacts of the differing format generation algorithm. For the other two scenarios, there are no clear biases towards verbal or numerical formats visible – the two scenarios have complementary performances of the two format groups (Fig. 4.15).

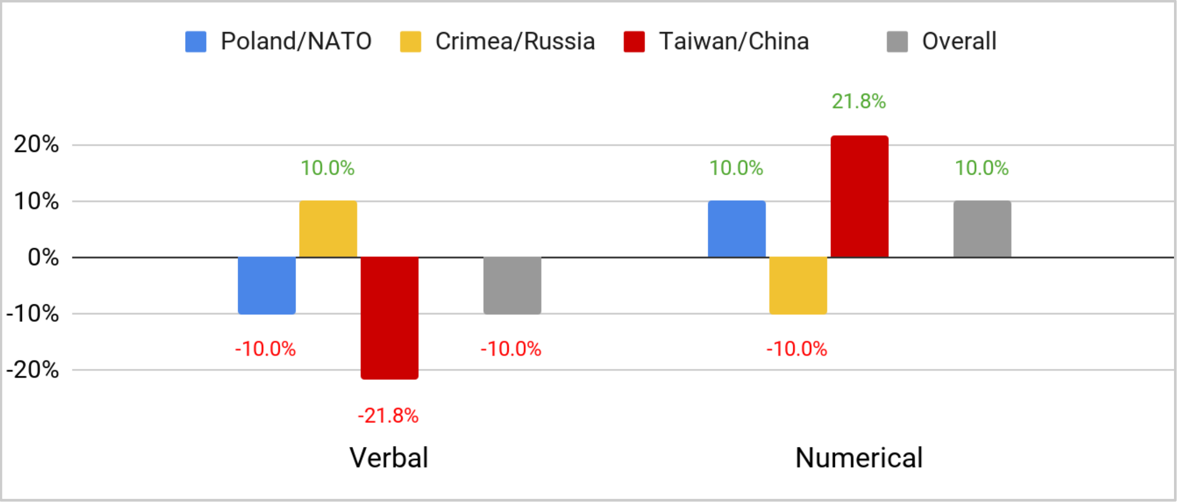


Figure 4.15: *Over- or underperformance of the verbal and numerical formats across scenarios (relative to the expected frequencies), expert cohort*

Besides the possible interference of the format generation algorithm, it should also be noted that the expert cohort is quite small compared to the non-expert cohort – especially when considering that each participant only got the task to select the 'expert assessments' in one scenario. Therefore, the few results of statistical significance from this cohort in the Taiwan/China scenario should be taken with a considerable grain of salt.

Overall, this analysis concludes that the data does not support the hypothesis that numerical formats are associated with a higher expertise of an assessment's producer.

Secondary analytical Aims

In the descriptive part of the data analysis, inconsistencies between the choice of VPE and the numeric probability estimate were detected. Particularly, a reduction in inconsistent outliers from the first to the second scenario was apparent. That could point to an acclimatisation effect – the participants having to 'get used to' expressing their assessment adequately numerically. The improvement in consistency from the first to the second scenario would then be the result of participants' learning curve. Alternatively, it could be the sign of differences in the intuitiveness, or 'ease of use', of the numeric format in different probability ranges. The first scenario was mostly assessed as "highly unlikely" or "unlikely". These VPEs have negative directionality (e.g. see Teigen/Brun, 1995). Numeric probabilities, on the other hand, have positive directionality (Teigen/Brun, 2000) e.g. even something "highly *un*likely" would have a positive likelihood of, say, 10%). In the second scenario, which was mostly assessed "highly likely" or "likely", the VPEs had positive directionality, just as the numeric probabilities.

Fig. 4.16 plots the share of inconsistent estimates across the whole probability range.[42] One caveat that has to be taken into account is the limitation that the survey offered no VPE option to express precisely even odds (there was nothing between "unlikely" and "likely"). Thus, numerical probability estimates of 50% would always be inconsistent with the VPE that had been chosen previously. For that reason, all inconsistent 50%-estimates (grey bar in Fig. 4.16) will be disregarded from further analysis.

[42] The frequencies are relative to the total number of probability estimates in a given likelihood interval.

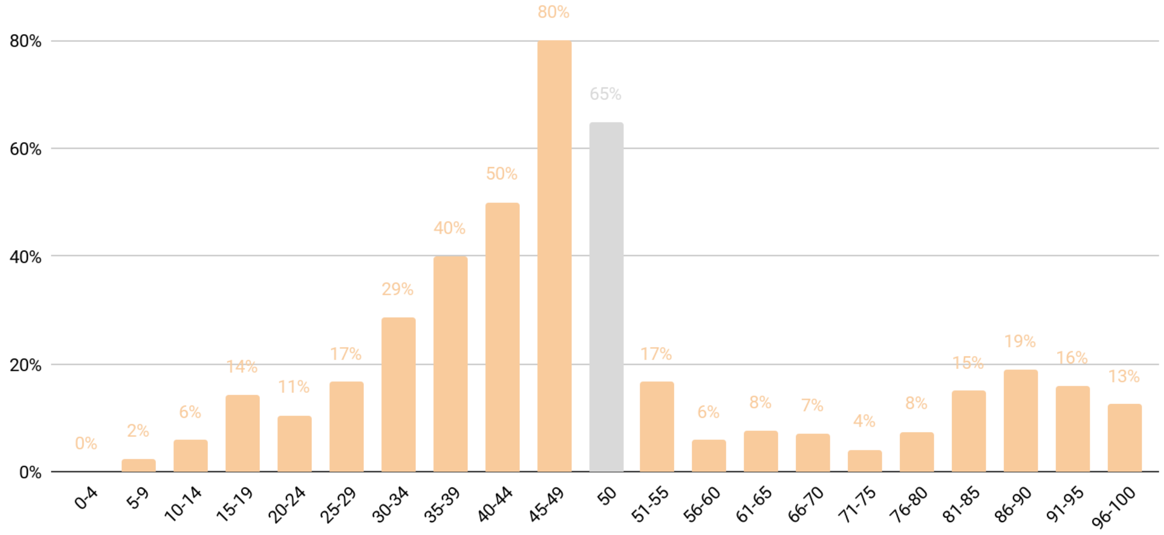


Figure 4.16: *Histogram of relative frequencies of inconsistent numeric likelihood estimates across the whole probability range (in percent).*

Below the 50% likelihood mark, the incidence of inconsistent estimates seems to pretty steadily increase. Beyond 50%, the relative frequencies of inconsistent estimates more or less look like random noise, rather than resembling some particular trend or pattern. In order to get a more comprehensive picture of inconsistency throughout the course of the survey, Table 4.15 lists the relative frequencies of inconsistent (non-50%) numeric estimates across scenarios and between cohorts and directionality.

Inconsistent estimates					
Poland/NATO		**Crimea/Russia**		**Taiwan/China**	
Overall:	21 (14%)	Overall:	6 (4%)	Overall:	21 (14%)
Expert cohort:	5 (15%)	Expert cohort:	0 (0%)	Expert cohort:	5 (15%)
Non-expert cohort:	16 (13%)	Non-expert cohort:	6 (5%)	Non-expert cohort:	16 (14%)
Negative directionality:	20 (14%)	Negative directionality:	1 (20%)	Negative directionality:	5 (13%)
Positive directionality:	1 (13%)	Positive directionality:	5 (3%)	Positive directionality:	16 (14%)

Table 4.15: *Absolute and relative (in brackets) frequencies of inconsistent estimates between cohorts and between directionalities (of the VPE) across scenarios*

This representation clearly shows that the acclimatisation hypothesis does not hold up. While the number of inconsistent numeric estimates does drop from scenario 1 to scenario 2, inconsistent estimates rise to their original levels in scenario 3. Across all three scenarios, 13.68% of verbal assessments with a negative directionality had an inconsistent numeric equivalent while for the VPEs with a positive directionality, 8.17% had an inconsistent numeric interpretation. To test whether that difference is statistically significant, a mixed effect logistic regression model is used again. Directionality is treated as the independent variable and the inconsistency as the dependent variable (both coded as binary variables), while accounting for the repeated measures by including Participant ID (each participant numbered from 1 to 153) as a random effect (Table 4.16).

Generalised linear mixed model fit by maximum likelihood (Laplace Approximation)

Overall dataset				
Predictor	**Estimate (β)**	**Std. Error**	***z*-value**	***p*-value**
(Intercept)	-2.1073	0.3299	-6.389	< .001***
Directionality	-0.6480	0.3332	-1.945	.0518
Random effects	**Variance**	**Std. Dev.**		
Participant ID	1.081	1.04		

Table 4.16: *Mixed effects logistic regression model for numeric estimate inconsistency*

The predictor directionality is only 0.18 percentage points shy of the 5% significance threshold (*p* = .0518). While that's technically not small enough to reject the null hypothesis at the conventionally set significance level, the effect of directionality on estimate inconsistency should still be considered. The parameter estimate (β_1 = -0.648) means that positive directionality halves the odds of a participant's numeric estimate turning out inconsistent (because of the odds ratio: $e^{-0.648} \approx 0.52309$). Or put differently: Negative directionality about doubles the chance of the numeric estimate being inconsistent.

Much of the data analysis so far – especially that related to hypothesis 1 and 2 – has focussed on changes in format preference. But throughout the different scenarios, levels of uncertainty and settings, there always remained considerable shares of participants that did not change preferences between phases (Fig. 4.17).

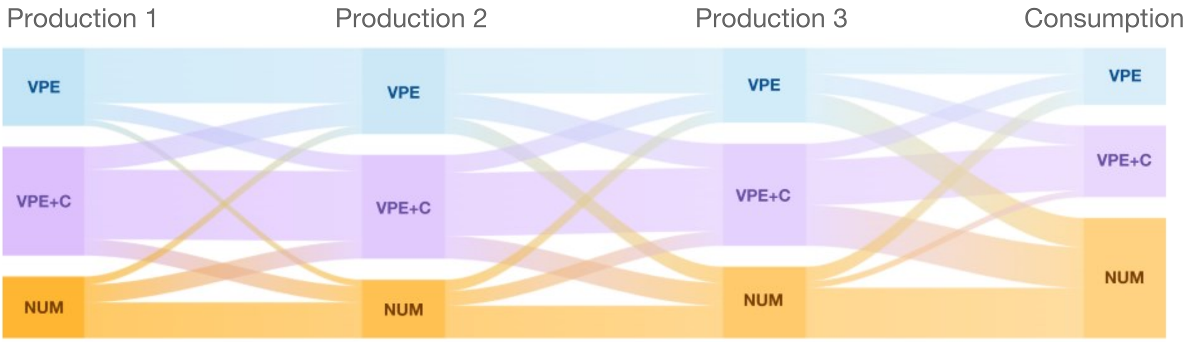


Figure 4.17: *Sankey diagram of format preferences throughout the experiment, all participants*

This final section of the data analysis seeks to explore this *preference stability*. For terminology, the analysis will borrow the two archetypes Sherman Kent coined: 'poets' and 'mathematicians' (which were already discussed in *1.4 Preferences and Politics*). Assuming that there are predetermined, underlying – and somewhat rigid – preferences, it is worth looking at the relative frequencies of format preference change versus preference stability in the production phase of the survey.

Tables 4.17.1 through 4.17.3 plot exactly that – the columns represent all possible combinations of preference change or non-change between assessment 1 (Poland/NATO) and assessment 2 (Crimea/Russia); the rows do the same but with regards to assessment 2 (Crimea/Russia) and assessment 3 (Taiwan/China). Verbal format preference is denoted as V and numerical format preference as N. Cells in the contingency table indicate the relative frequency of the given preference combination.

N = 153	$V_1 \to V_2$	$N_1 \to V_2$	$V_1 \to N_2$	$N_1 \to N_2$
$V_2 \to V_3$	50.59%	7.19%		
$N_2 \to V_3$			5.88%	4.58%
$V_2 \to N_3$	12.42%	3.27%		
$N_2 \to N_3$			3.27%	9.8%

- 🟨 poets (50.59%)
- 🟩 ⅔ poet, ⅓ mathem. (25.49%)
- ▨ ⅔ mathem., ⅓ poet (11.12%)
- 🟦 mathematicians (9.8%)

Table 4.17.1: *Contingency table plotting relative frequencies of format change or stability, all participants*

n = 119	$V_1 \to V_2$	$N_1 \to V_2$	$V_1 \to N_2$	$N_1 \to N_2$
$V_2 \to V_3$	50.42%	8.4%		
$N_2 \to V_3$			6.72%	5.04%
$V_2 \to N_3$	10.08%	3.36%		
$N_2 \to N_3$			4.2%	11.76%

- 🟨 poets (50.42%)
- 🟩 ⅔ poet, ⅓ mathem. (25.2%)
- ▨ ⅔ mathem., ⅓ poet (12.6%)
- 🟦 mathematicians (11.76%)

Table 4.17.2: *Contingency table plotting relative frequencies of format change or stability, non-expert cohort*

n = 34	$V_1 \to V_2$	$N_1 \to V_2$	$V_1 \to N_2$	$N_1 \to N_2$
$V_2 \to V_3$	64.71%	2.94%		
$N_2 \to V_3$			2.94%	17.65%
$V_2 \to N_3$	20.59%	2.94%		
$N_2 \to N_3$			0%	2.94%

- 🟨 poets (64.71%)
- 🟩 ⅔ poet, ⅓ mathem. (26.47%)
- ▨ ⅔ mathem., ⅓ poet (20.59%)
- 🟦 mathematicians (2.94%)

Table 4.17.3: *Contingency table plotting relative frequencies of format change or stability, expert cohort*

About half (50.4%) of all participants are *poets* – i.e. they strictly preferred a verbal format for each assessment they produced. For the expert cohort, this share is even higher (64.7%). This delta might be explained by the fact that about half of the expert participants (52.9%) reported using (non-standardised) verbal probability expressions in their professional life; 35.3% even use specific NBLP schemes. This professional experience might cause a certain bias towards verbal formats: Two thirds (66.7%) of *poets* in the expert cohort also use NBLP schemes in their real-life work.

Mathematicians, on the other hand, are very rare in the expert cohort (only 1 in 34, amounting to 2.9%); whereas about every tenth participant in the non-expert cohort (11.8%) stuck to submitting numerical assessments.

In the mixed groups (i.e. participants who chose both numerical and verbal formats throughout the estimate production phase), the expert cohort stands out. A considerable share (17.7%) of expert participants generally favoured the numerical format but changed their choice to a verbal format in the last scenario. That behaviour is in line with what the congruence principle would expect: For the first two scenarios, the precision of the numerical format was preferred; but when it came to the Taiwan/China scenario – the scenario with the highest uncertainty by most measures discussed previously[43] – participants pivoted to a more vague verbal format to allow for ambiguity. In the analysis of hypothesis 2, however, this behaviour didn't show, because it was overshadowed by an even larger share of expert participants (20.6%) who behaved exactly contrary – opting for verbal formats in the first two scenarios but then choosing the numerical format for the Taiwan/China assessment.

This dynamic, which precisely runs counter to the congruence principle, can also be seen in the consumer setting – and again, it is only present in the expert cohort (Fig 4.18). For the Poland/NATO scenario, which contained, arguably, the least uncertainty, experts overwhelmingly preferred verbal formats. For the Taiwan/China scenario – the most uncertain one – over 70% favoured the numerical format. Also, no expert preferred a 'simple' verbal assessment just stating the likelihood.

[43] This scenario's median level of analytic confidence is lower than for the first two scenarios (Fig. 4.6.1-4.6.3), its mean likelihood is the further away from the extremes 0 or 100% (Table 4.4.1) and its mean error margin is the largest across all three scenarios (Table 4.4.2).

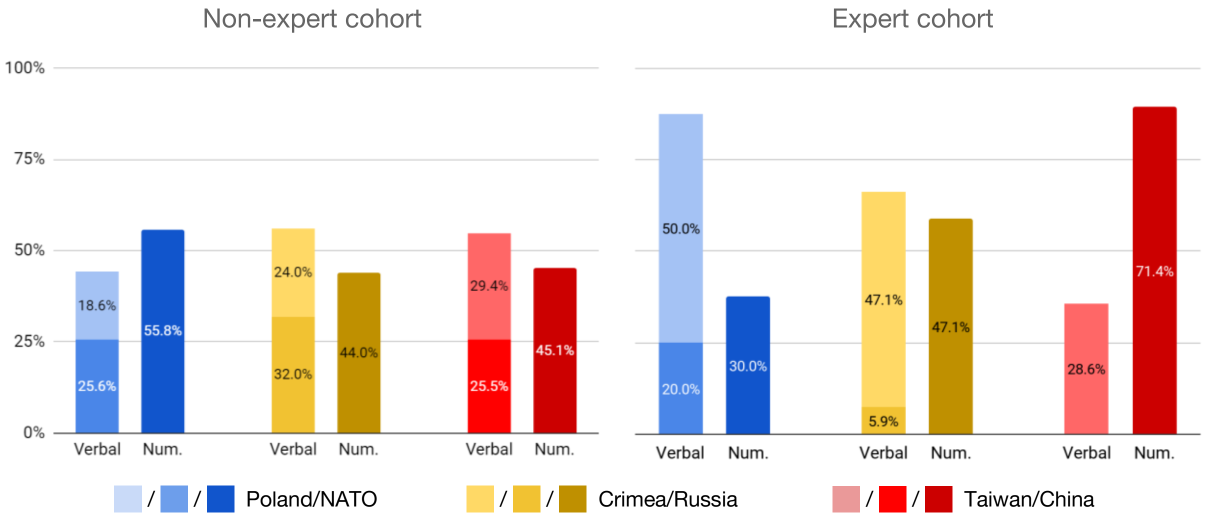


Figure 4.18: *Format preferences of non-experts (left) and experts (right) grouped by scenario*

As discussed earlier (in *Hypothesis 2 – Congruence Principle*), this effect of the scenario on format preference in the expert cohort falls short of statistical significance, but not by much (*p* = .05971). And since the 0.05 significance threshold is more of an agreed-upon convention than a sacrosanct hard border, this result is worth being taken seriously. It suggests that, both in the production and in the consumption setting, there is a certain demand for more information about the uncertainties of an assessment, the more uncertain the assessed matter is. For more certain assessments, on the other hand, more vagueness and imprecision – i.e. less information about uncertainty – is perfectly acceptable or even preferred.

To conclude the data analysis, a last point on preference stability: As discussed at length in previous sections of the analysis, a change in setting (from producer to consumer) significantly affects the overall preference distribution; with a significant number of participants shifting from preferring the verbal to the numerical format. Still, a considerable number of participants (58.8%) stayed with their initial format preference from the production phase. Figure 4.19 plots the conditional probabilities of the format preference either being verbal or numerical, given that format preference remained stable in all previous assessments. The starting point (left side) is the initially observed probability of any participant choosing either a verbal P(V_1) or a numerical format P(N_1) in the first assessment. The graph then goes through the production phase, plotting the conditional probabilities of format stability at each assessment. Lastly, the graph plots the probability that format preference remained stable into the consumption setting, given that the format preference was never changed in the production setting. What this graph essentially shows is: How likely is a *poet* or a *mathematician* to stay a *poet* or a *mathematician* in the next task of the experiment?

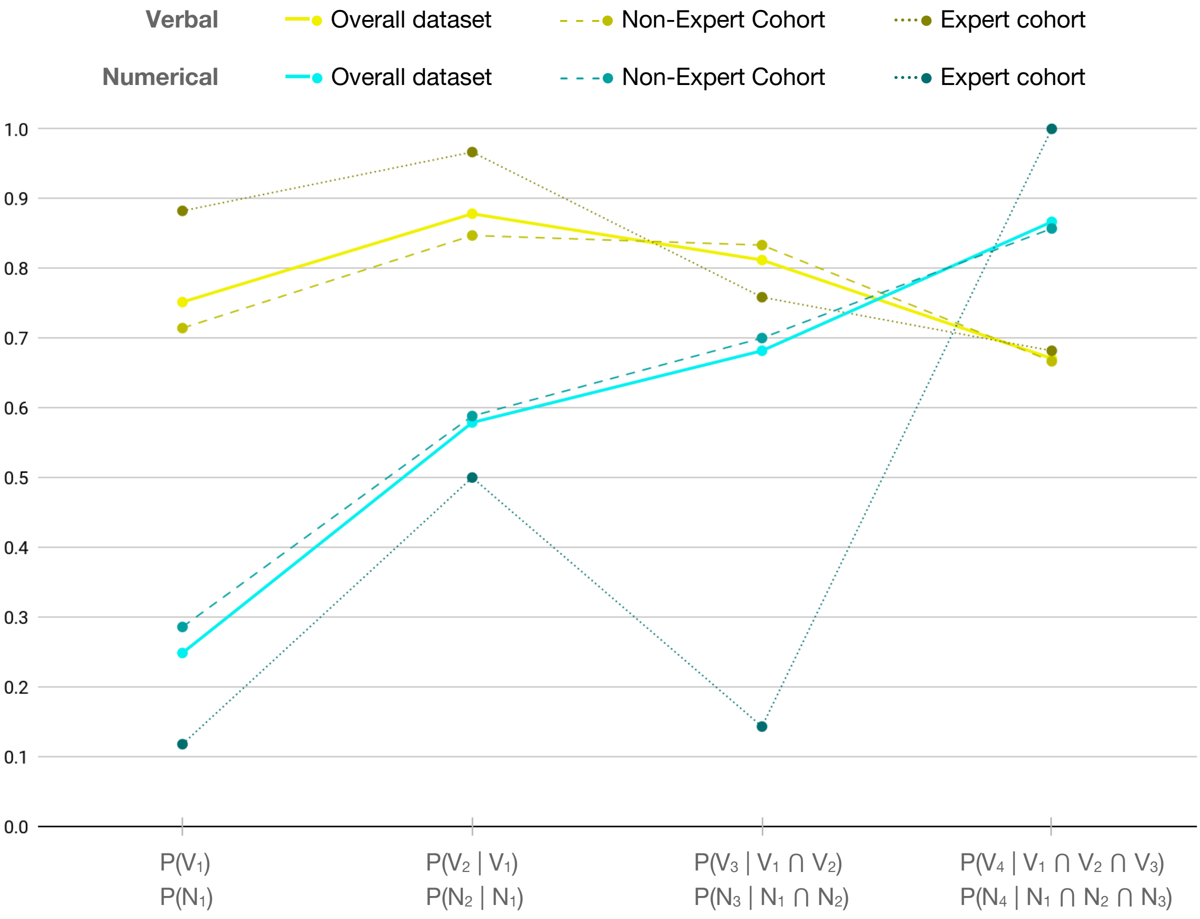


Figure 4.19: *Conditional probabilities of format preference stability*

First of all, the initial incidence of *mathematicians* (P(V_1) = 0.2484; *n* = 38) is much lower than that of the *poets* (P(N_1) = 0.7516; *n* = 115). But their preference stability quickly increases throughout the survey – already at the second assessment, it is more likely that a participant chose a numerical format, given a numerical preference in the first assessment. At the end of the survey, in the consumption phase, the probability of *mathematicians*' preference stability reaches its maximum $P(N_4 \mid N_1 \cap N_2 \cap N_3) = 0.8667$ ($n = 13$). For the *poets*, preference stability does not increase throughout the course of the survey, despite being quite high from the get go (particularly compared to the *mathematicians*). In the production phase, the *poets* were very likely not to change their format preference ($P(V_2 \mid V_1) = 0.8783$ ($n = 101$) and $P(V_3 \mid V_1 \cap V_2) = 0.8119$ ($n = 82$). Going into the consumption setting, however, *poets* are affected by the effect discussed in the context of hypothesis 1: conditional format stability drops to $P(V_4 \mid V_1 \cap V_2 \cap V_3) = 0.6707$ ($n = 55$).

In summary, the analysis on preference stability highlights three key aspects: First, *poets* (50.59%) outnumber *mathematicians* (9.8%) five to one – among the experts even 22:1 (60.71% versus 2.94%), possibly due to professional habits. Because of the dynamics addressed with regards to hypothesis 1, *poets* are vulnerable to preference instability when the setting is changed. But still, and that is the second key point, two thirds of *poets* (67%) maintain preference stability even under that condition. Lastly, *mathematicians*' preference stability – while staying below *poets*' throughout assessment production – profited off of the change in setting (from production to consumption), surpassing *poets*' format stability.

5. General Discussion of the Empirical Findings

In order to put the panoply of empirical findings from the data analysis into perspective, this section will start out by addressing the three key analytical questions and put them into the context of the existing research. That is followed up by a reflection upon the additional findings of this research.

1. Do Estimate Producers prefer Words while Consumers prefer Numbers?

While most producers did prefer VPEs, consumer preferences were pretty evenly split between verbal and numerical formats. Irwin and Mandel (2023) also found an even split in format preference – however, only with their expert cohort of 41 Canadian intelligence analysts. The majority of their non-expert cohort (440 individuals from Canada and the US, recruited over an online crowdsourcing service) did, in fact, favour numbers over words in the consumer setting (Irwin/Mandel, 2023). The non-expert cohort in this research displayed a rather similar behaviour to the expert participants – with verbal and numerical formats head to head in the consumer preferences. The reason for this difference in non-expert behaviour between Irwin and Mandel's and this research might be the fact that many non-experts in this analysis were still quite elite participants. Only those who reported to be "often" the consumer or producer of estimative judgements qualified for the expert cohort – thus, a considerable number of participants that "sometimes" or "rarely" professionally consumed or produced estimative judgements on national security did end up in the non-expert cohort. Additionally, due to resource and recruitment limitations, most participants of this research have some background in International Relations, Security Studies, public service or the military.

However, format preferences did see a significant shift between the producer and the consumer setting in favour of the numeric format.

2. Does the Congruence Principle play a Role in this Context?

Only a small subset of participants acted in accordance with the congruence principle – i.e. changed their format preference from verbal for more uncertain assessments to numerical for assessments of higher certainty. Overall, no evidence for the congruence principle being a factor was found. In fact, the data shows significant evidence for the exact opposite behaviour. Possible implications of that finding will be discussed in the conclusion.

3. Does the Risk of an Illusion of Rigour exist in this Context?

The data does not show significant evidence that numeric probabilities evoked a false sense of expertise. This result fits pretty neatly into the existing literature, which conceptualises trustworthiness and expertise as the two key components of credibility (Hovland et al., 1953; Wiener/Mowen, 1986). Friedman et al. (2017) found that probabilistic information in a security-related context was not trusted more when conveyed in a numerical format instead of a verbal format. Jenkins et al.'s (2017) research found no format effect on overall credibility. This research's finding of no association between format and perceived expertise hence fits into this theoretical equation of credibility being a function of trust and expertise.[44]

While some concerns about the use of numeric probabilities in estimative judgements that relate to the consumer side – e.g. like the 'illusion of rigour' – are not supported by this research, there seem to be some challenges on the producer side.

[44] Yet, this notion that using numeric probabilities created some sense of unwarranted 'pseudo-certainty' of estimative assessments is still a very prevalent concern – which even some participants of this survey brought up (in post-survey interaction or via the free-text input in phase 4 of the experiment).

4. Inconsistencies – Numeric Likelihood

Some participants did not produce consistent numeric likelihood estimates. The analysis has shown that such inconsistencies were twice as likely when the previously made verbal likelihood assessment had negative directionality. In recent years, much research has investigated the importance of directionality in uncertainty communication, but mostly from the receiver's side. For instance, Collins et al. (2023) found that directionality affects the interpretation of probabilistic assessments and thereby influences decision-making – particularly in low probability conditions, where numerical formats and verbal formats diverge in their directionality. This research complements the existing canon by showing that similar dynamics also seem to play out in the production of probabilistic estimates.

Psychological biases have taken centre stage in thinking about intelligence analysis over the past three decades (e.g. Heuer, 1999). They have been addressed, for example, through the development of Structured Analytic Techniques (Heuer/Pherson, 2019) to mitigate the risks of 'system 1 thinking' (Kahneman, 2011; Pherson et al., 2024). With regards to subjective probabilities in intelligence assessments, directionality is certainly a factor to be cognisant of.

5. Inconsistencies – Analytic Confidence

Participants' use of the concept of analytic confidence was not always consistent. First of all, participants rarely used the expression "low analytic confidence". To express limited confidence, participants mostly opted for the expression "moderate confidence" – even in the third scenario which had a time horizon of 26 years and in which participants' likelihood assessments were spread all over the probability range. That could either be a sign of participants' overconfidence or of bad calibration in the application of the concept. Overconfidence has been identified both in accomplished geopolitical experts (Tetlock, 2005) as well as in medical diagnoses or legal judgements (Dhami et al., 2015). In contrast to some of these findings, Mandel and Barnes (2014) found that strategic intelligence forecasts rather suffered from underconfidence.

This research design was not geared towards finding a definite answer on whether participants were overconfident or not. What can be concluded, however, is that numeric interpretations of the three confidence levels were inconsistent. No matter if the verbal confidence level was “low” or “moderate”, the assigned error margins mostly landed around 10 percentage points. Again, this research is not able to disambiguate whether this is caused by overconfidence or if it is just a sign of non-linear interpretation of these confidence levels – or if both are at play. Both phenomena would not be unexpected. Friedman et al. (2016) found that participants were particularly overconfident in numeric estimates. Duke (2024) found that many participants did not associate higher confidence with narrower probability ranges – which could be the reason behind the non-linear numeric interpretation found in this experiment. Although, participants were able to discriminate between “high” and “moderate” confidence when determining the size of the error margin.

On how well the numeric range format is handled by receivers of such estimates is also not conclusively determined. While Irwin and Mandel (2023) found that participants understood the width of the interval as intended – the wider, the lower the analytic confidence –, Duke’s (2024) research found that participants did not interpret confidence as the uncertainty around likelihood estimate

Conclusion

Based on this research's results, and in line with existing academic literature, fears about numeric probabilities evoking a false sense of expertise are not warranted. Instead, there is demand for numeric precision – particularly, when certainty is low. The research found that even producers of probabilistic assessments tend to provide precise, numeric information in more uncertain assessments while using more ambiguous, verbal formats assessments with higher analytic confidence.

That suggests that there is not as much need to replace VPEs like "highly likely" or "highly unlikely" with 90% or 10% probability estimates. Where clarification is needed is in the middle of the range (e.g. between 20% and 80%) – there, numeric precision really adds informational value: to disambiguate whether "likely" means 60% or 80% likelihood[45] and, through the use of probability ranges, to specify whether "unlikely" with "low confidence" means 10%-40% or if it means 30%-60%.

Recommendations

Rather than sweeping universal reform to improve uncertainty communication, maybe a more 'strategic' use of numeric probabilities – only where VPEs are too ambiguous – would be a good way forward. Given the reluctance of intelligence communities to abandon VPEs, such a 'less invasive' approach that only seeks to selectively add numerical clarity where it is most needed, might be the most feasible.

This research has shown that there are some challenges on the producer side, namely in terms of calibrating the width of the interval to adequately convey analytic confidence. If intelligence communities[46] want to adopt numerical quantifiers in the form of probability ranges, attention should be paid to analysts' calibration in

[45] In the current NATO standard, "likely" can mean both (Mandel/Irwin, 2021: 560).

[46] or any other organisation involved in assessing the uncertainties of foreign and security policy.

adequately using the width of the interval to convey analytic confidence. One way to go about that could be to define specific error margins or interval widths as anchor points – as sort of a 'reversed' NBLP scheme, defining what interval widths correspond to what level of analytic confidence.

Further Research

The issue of improving the communication of second order probability to decision-makers still needs further research. Such research is important since there are plenty of shortcomings of the current system – it is imprecise, and it gets confused and conflated with first order probability by both intelligence consumers and producers. But it is important information for a decision-maker to receive, in order to put an analyst's assessment into context.

A related problem to be further looked into is the challenge of asymmetrical probability distributions. In much of the literature, as well as in this research, numeric likelihoods are treated as means of symmetrical probability distributions – i.e. a range estimate of 50% to 70% likelihood is equivalent to 60% plus or minus 10 percentage points. But it is entirely imaginable that, in some assessments, the underlying probability distribution might be skewed to one side. For example, the likelihood of regime change in a foreign country due to civil unrest might be assessed as being at 60% – but that the chance of the protest movement failing was higher than it exceeding expectations. In such a case, 60% would represent the mode rather than the mean with the range estimate being 45% to 65%.

Looking ahead, advancements in the utilisation of machine learning, specifically large language models and generative AI, in intelligence analysis will open up new debates and revitalise old ones. For instance, already today, ChatGPT is perfectly able to write a comprehensive assessment analysing the prospects of China attacking Taiwan by the middle of this century – listing various factors to consider and weighing them against each other to finally arrive at likelihood assessments for the short, medium and long term. Since such AI models work probabilistically, that opens up new possibilities for generating numeric probability estimates – why have a machine that thinks in precise probabilities and then limit its output to vague VPEs? On the other hand, the use of AI tools will raise concerns about credibility, 'illusions of rigour' and how decision-makers would generally perceive AI-assisted or even AI-generated analysis products.

Finally, one last general point: It seems that, if Intelligence Studies were a map, the United States (and maybe Canada and the UK) would be mapped fairly detailedly – while the rest of the world would still be rather blank, with only rudimentary borders drawn. That is a fundamental limitation that also this research suffers from. Much of the debates around and research on intelligence analysis and its interaction with decision-making stem from the Anglo-Saxon or, in many cases, from the American sphere. Even in fellow NATO countries in continental Europe, research on intelligence as a process and as a product is still underdeveloped. Presumably, not because of a lack of academic interest, but due to limited access to data and information. At least theoretically, there is plenty of potential for research on intelligence and intelligence-related matters outside the Anglo-Saxon context. In order to advance existing practices, to improve the products that are put out and maybe even to develop alternatives to Anglo-Saxon paradigms, more openness to and engagement with academia might be beneficial for European intelligence communities.

List of References

Agresti, A. (2007). *An introduction to categorical data analysis* (2nd ed). Wiley-Interscience.

Barnes, A. (2016). Making Intelligence Analysis More Intelligent: Using Numeric Probabilities. *Intelligence and National Security, 31*(3), 327–344. https://doi.org/10.1080/02684527.2014.994955

Bates, S., & Rosenbloom, J. L. (1998). *Kennedy and the Bay of Pigs* (Nos. C14-80–279; Kennedy School of Government Case Program, p. 31). Harvard Kennedy School.

Bergenstrom, A., & Sherr, L. (2003). The effect of order of presentation of verbal probability expressions on numerical estimates in a medical context. *Psychology, Health & Medicine, 8*(4), 391–398. https://doi.org/10.1080/1354850310001604522

Brun, W., & Teigen, K. H. (1988). Verbal probabilities: Ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes, 41*(3), 390–404. https://doi.org/10.1016/0749-5978(88)90036-2

Budescu, D. V., & Wallsten, T. S. (1985). Consistency in interpretation of probabilistic phrases. *Organizational Behavior and Human Decision Processes, 36*(3), 391–405. https://doi.org/10.1016/0749-5978(85)90007-X

Budescu, D. V., & Wallsten, T. S. (1995). Processing Linguistic Probabilities: General Principles and Empirical Evidence. In *Psychology of Learning and Motivation* (Vol. 32, pp. 275–318). Elsevier. https://doi.org/10.1016/S0079-7421(08)60313-8

Budescu, D. V., Weinberg, S., & Wallsten, T. S. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance, 14*(2), 281–294. https://doi.org/10.1037/0096-1523.14.2.281

CIA. (1951). NIE-29: PROBABILITY OF AN INVASION OF YUGOSLAVIA IN 1951 (No. 29). CIA; CREST (CIA Records Search Tool). https://www.cia.gov/readingroom/docs/CIA-RDP79R01012A000700040018-0.pdf

Cochran, W. G. (1954). Some Methods for Strengthening the Common χ 2 Tests. *Biometrics, 10*(4), 417. https://doi.org/10.2307/3001616

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed). L. Erlbaum Associates.

Cohn, L. (2009). Quantifying Risk: Verbal Probability Expressions in Spanish and English. *American Journal of Health Behavior, 33*(3). https://doi.org/10.5993/AJHB.33.3.3

Collins, R. N., Mandel, D. R., & MacLeod, B. A. (2023). Verbal and numeric probabilities differentially shape decisions. *Thinking & Reasoning*, 1–23. https://doi.org/10.1080/13546783.2023.2220971

Dahl, E. J. (2013). Why Won't They Listen? Comparing Receptivity Toward Intelligence at Pearl Harbor and Midway. *Intelligence and National Security, 28*(1), 68–90. https://doi.org/10.1080/02684527.2012.749061

Damrosch, S. P., & Soeken, K. (1983). Communicating probability in clinical reports: Nurses' numerical associations to verbal expressions. *Research in Nursing & Health, 6*(2), 85–87. https://doi.org/10.1002/nur.4770060208

Davis, J. (1991). The Kent-Kendall Debate of 1949. *Studies in Intelligence, 35*(2), 37–50. https://www.cia.gov/static/Kent-Kendall-Debate-1949.pdf

Defence Intelligence. (2023, February 17). *Defence Intelligence – communicating probability*. Gov.uk. https://www.gov.uk/government/news/defence-intelligence-communicating-probability

Dhami, M. K. (2018). Towards an evidence-based approach to communicating uncertainty in intelligence analysis. *Intelligence and National Security, 33*(2), 257–272. https://doi.org/10.1080/02684527.2017.1394252

Dhami, M. K., & Mandel, D. R. (2021). Words or numbers? Communicating probability in intelligence analysis. *American Psychologist, 76*(3), 549–560. https://doi.org/10.1037/amp0000637

Dhami, M. K., & Mandel, D. R. (2022). Communicating uncertainty using words and numbers. *Trends in Cognitive Sciences, 26*(6), 514–526. https://doi.org/10.1016/j.tics.2022.03.002

Dhami, M. K., Mandel, D. R., Mellers, B. A., & Tetlock, P. E. (2015). Improving Intelligence Analysis With Decision Science. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science, 10*(6), 753–757. https://doi.org/10.1177/1745691615598511

Dhami, M. K., & Wallsten, T. S. (2005). Interpersonal comparison of subjective probabilities: Toward translating linguistic probabilities. *Memory & Cognition, 33*(6), 1057–1068. https://doi.org/10.3758/BF03193213

DIA. (1976). *Military Significance of Soviet Developed Facilities in Somalia (U)* (Defense Intelligence Estimate No. DIE SOV 2-76; p. 18). Defense Intelligence Agency Directorate for Estimates. https://2001-2009.state.gov/documents/organization/67018.pdf

Dieckmann, N. F., Mauro, R., & Slovic, P. (2010). The Effects of Presenting Imprecise Probabilities in Intelligence Forecasts: Effects of Presenting Imprecise Probabilities in Intelligence Forecasts. *Risk Analysis, 30*(6), 987–1001. https://doi.org/10.1111/j.1539-6924.2010.01384.x

Duke, M. C. (2024). Probability and confidence: How to improve communication of uncertainty about uncertainty in intelligence analysis. *Journal of Behavioral Decision Making, 37*(1), e2364. https://doi.org/10.1002/bdm.2364

Edwards, A. L. (1948). Note on the “correction for continuity” in testing the significance of the difference between correlated proportions. *Psychometrika, 13*(3), 185–187. https://doi.org/10.1007/BF02289261

Friedman, J. A. (2019). *War and chance: assessing uncertainty in international politics*. Oxford University Press.

Friedman, J. A., Baker, J. D., Mellers, B. A., Tetlock, P. E., & Zeckhauser, R. (2018). The Value of Precision in Probability Assessment: Evidence from a Large-Scale Geopolitical Forecasting Tournament. *International Studies Quarterly*. https://doi.org/10.1093/isq/sqx078

Friedman, J. A., Lerner, J. S., & Zeckhauser, R. (2017). Behavioral Consequences of Probabilistic Precision: Experimental Evidence from National Security Professionals. *International Organization, 71*(4), 803–826. https://doi.org/10.1017/S0020818317000352

Friedman, J. A., & Zeckhauser, R. (2015). Handling and Mishandling Estimative Probability: Likelihood, Confidence, and the Search for Bin Laden. *Intelligence and National Security, 30*(1), 77–99. https://doi.org/10.1080/02684527.2014.885202

Friedman, J. A., & Zeckhauser, R. (2018). Analytic Confidence and Political Decision-Making: Theoretical Principles and Experimental Evidence From National Security Professionals: Analytic Confidence and Political Decision-Making. *Political Psychology, 39*(5), 1069–1087. https://doi.org/10.1111/pops.12465

Friedman, J., Lerner, J., & Zeckhauser, R. J. (2016). How Quantifying Probability Assessments Influences Analysis and Decision Making: Experimental Evidence from National Security Professionals. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2782598

Friedman, J., & Zeckhauser, R. J. (2012). *Assessing Uncertainty in Intelligence* (SSRN Scholarly Paper No. 2087604). https://doi.org/10.2139/ssrn.2087604

Gates, R. M. (1992). *Guarding against politicization*. https://doi.apa.org/doi/10.1037/e741302011-003

Gaub, F. (2022). Fearsight: von der Furcht, konkret zu werden. In K. Schäfer, K. Steinmüller, & A. Zweck (Eds.), *Gefühlte Zukunft* (pp. 337–343). Springer Fachmedien Wiesbaden. https://doi.org/10.1007/978-3-658-35890-7_16

Gill, P. (2020). Explaining Intelligence Failure: Rethinking the Recent Terrorist Attacks in Europe. *International Journal of Intelligence and CounterIntelligence, 33*(1), 43–67. https://doi.org/10.1080/08850607.2019.1663702

Gill, P., & Phythian, M. (2013). From Intelligence Cycle to web of intelligence. In M. Phythian (Ed.), *Understanding the Intelligence Cycle*. Routledge.

Gurmankin, A. D., Baron, J., & Armstrong, K. (2004). The Effect of Numerical Statements of Risk on Trust and Comfort with Hypothetical Physician Risk Communication. *Medical Decision Making, 24*(3), 265–271. https://doi.org/10.1177/0272989X04265482

Hamm, R. M. (1991). Selection of verbal probabilities: A solution for some problems of verbal probability expression. *Organizational Behavior and Human Decision Processes, 48*(2), 193–223. https://doi.org/10.1016/0749-5978(91)90012-I

Han, P. K. J., Klein, W. M. P., Lehman, T. C., Massett, H., Lee, S. C., & Freedman, A. N. (2009). Laypersons' responses to the communication of uncertainty regarding cancer risk estimates. *Medical Decision Making: An International Journal of the Society for Medical Decision Making, 29*(3), 391–403. https://doi.org/10.1177/0272989X08327396

Heuer, R. J. (1999). *Psychology of intelligence analysis*. Center for the Study of Intelligence, Central Intelligence Agency.

Ho, E. H., Budescu, D. V., Dhami, M. K., & Mandel, D. R. (2015). Improving the communication of uncertainty in climate science and intelligence analysis. *Behavioral Science & Policy, 1*(2), 43–55. https://doi.org/10.1353/bsp.2015.0015

Hobby, J. L., Tom, B. D., Todd, C., Bearcroft, P. W., & Dixon, A. K. (2000). Communication of doubt and certainty in radiological reports. The British *Journal of Radiology, 73*(873), 999–1001. https://doi.org/10.1259/bjr.73.873.11064655

Holtgraves, T., & Perdew, A. (2016). Politeness and the communication of uncertainty. *Cognition, 154*, 1–10. https://doi.org/10.1016/j.cognition.2016.05.005

Honda, H., & Yamagishi, K. (2006). Directional Verbal Probabilities: Inconsistencies Between Preferential Judgments and Numerical Meanings. *Experimental Psychology, 53*(3), 161–170. https://doi.org/10.1027/1618-3169.53.3.161

Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion*. Yale University Press.

Irwin, D., & Mandel, D. (2019). *Variants of Vague Verbiage: Intelligence Community Methods for Communicating Probability* (SSRN Scholarly Paper No. 3441269). https://doi.org/10.2139/ssrn.3441269

Irwin, D., & Mandel, D. R. (2023). Communicating uncertainty in national security intelligence: Expert and nonexpert interpretations of and preferences for verbal and numeric formats. *Risk Analysis, 43*(5), 943–957. https://doi.org/10.1111/risa.14009

Jenkins, S. C., Harris, A. J. L., & Lark, R. M. (2017). *Maintaining credibility when communicating uncertainty: The role of communication format*. 582–587.

Jenkins, S. C., Harris, A. J. L., & Lark, R. M. (2018). Understanding 'Unlikely (20% Likelihood)' or '20% Likelihood (Unlikely)' Outcomes: The Robustness of the Extremity Effect. *Journal of Behavioral Decision Making, 31*(4), 572–586. https://doi.org/10.1002/bdm.2072

Jenkins, S. C., Harris, A. J. L., & Lark, R. M. (2019). When unlikely outcomes occur: the role of communication format in maintaining communicator credibility. *Journal of Risk Research, 22*(5), 537–554. https://doi.org/10.1080/13669877.2018.1440415

Jervis, R. (1976). *Perception and Misperception in International Politics*. Princeton University Press. https://doi.org/10.1515/9781400885114

Jervis, R. (1978). Cooperation under the Security Dilemma. *World Politics, 30*(2), 167–214. https://doi.org/10.2307/2009958

Jervis, R. (2010). *Why intelligence fails: lessons from the Iranian Revolution and the Iraq War*. Cornell University Press.

Joint Chiefs of Staff. (1961). *Memorandum From the Joint Chiefs of Staff to Secretary of Defense McNamara*. US Department of State Office of the Historian. https://history.state.gov/historicaldocuments/frus1961-63v10/d35

Joint Chiefs of Staff. (2013). *Joint Intelligence: Joint Publication 2-0*. https://irp.fas.org/doddir/dod/jp2_0.pdf

Juanchich, M., & Sirota, M. (2013). Do people really say it is “likely” when they believe it is only “possible”? Effect of politeness on risk communication. *Quarterly Journal of Experimental Psychology, 66*(7), 1268–1275. https://doi.org/10.1080/17470218.2013.804582

Kahneman, D. (2011). *Thinking, fast and slow* (1st ed). Farrar, Straus and Giroux.

Karelitz, T. M., & Budescu, D. V. (2004). You Say ‘Probable’ and I Say ‘Likely’: Improving Interpersonal Communication With Verbal Probability Phrases. *Journal of Experimental Psychology: Applied, 10*(1), 25–41. https://doi.org/10.1037/1076-898X.10.1.25

Kent, S. (1949). *Strategic Intelligence for American World Policy*. Princeton University Press. https://www.jstor.org/stable/j.ctt183q0qt

Kent, S. (1964). Words of Estimative Probability. *Studies in Intelligence, 8*(4), 49–65. https://www.cia.gov/readingroom/docs/CIA-RDP93T01132R000100020036-3.pdf

Kesselman, R. F. (2008). *Verbal Probability Expressions in National Intelligence Estimates: A Comprehensive Analysis of Trends from the Fifties through Post 9/11* [Master’s Thesis, Mercyhurst College]. https://gwern.net/doc/statistics/bayes/2008-kesselman.pdf

Kong, A., Barnett, G. O., Mosteller, F., & Youtz, C. (1986). How Medical Professionals Evaluate Expressions of Probability. *New England Journal of Medicine, 315*(12), 740–744. https://doi.org/10.1056/NEJM198609183151206

Lichtenstein, S., & Newman, J. R. (1967). Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychonomic Science, 9*(10), 563–564. https://doi.org/10.3758/BF03327890

Longman, T., Turner, R. M., King, M., & McCaffery, K. J. (2012). The effects of communicating uncertainty in quantitative health risk estimates. *Patient Education and Counseling, 89*(2), 252–259. https://doi.org/10.1016/j.pec.2012.07.010

Lowenthal, M. M. (1992). Tribal tongues: Intelligence consumers, intelligence producers. *The Washington Quarterly, 15*(1), 157–168. https://doi.org/10.1080/01636609209550084

Lowenthal, M. M. (2006). *Intelligence: from secrets to policy* (3rd ed). CQ Press.

Mandel, D. R. (2015). Accuracy of Intelligence Forecasts From the Intelligence Consumer's Perspective. *Policy Insights from the Behavioral and Brain Sciences, 2*(1), 111–120. https://doi.org/10.1177/2372732215602907

Mandel, D. R., & Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences of the United States of America, 111*(30), 10984–10989. https://doi.org/10.1073/pnas.1406138111

Mandel, D. R., & Irwin, D. (2021). Uncertainty, Intelligence, and National Security Decisionmaking. *International Journal of Intelligence and CounterIntelligence, 34*(3), 558–582. https://doi.org/10.1080/08850607.2020.1809056

Manjikian, M. (2020). "Those Clowns Out at Langley": A Theory of Trust between the Intelligence Community and the President. *International Journal of Intelligence and CounterIntelligence, 33*(4), 709–730. https://doi.org/10.1080/08850607.2020.1780084

Mantel, N., & Haenszel, W. (1959). Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. *JNCI: Journal of the National Cancer Institute*. https://doi.org/10.1093/jnci/22.4.719

Marchio, J. (2014). "If the Weatherman Can...": The Intelligence Community's Struggle to Express Analytic Uncertainty in the 1970s. *Studies in Intelligence, 58*(4), 31–42. https://www.cia.gov/static/673d0414fdd6f124560185cca4a61d26/ICs-Struggle-to-Express.pdf

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika, 12*(2), 153–157. https://doi.org/10.1007/BF02295996

Mearsheimer, J. J. (1994). The False Promise of International Institutions. *International Security, 19*(3), 5. https://doi.org/10.2307/2539078

NATO STO. (2020). *Assessment and Communication of Uncertainty in Intelligence to Support Decision-Making* (No. STO-TR-SAS-114; Technical Report RDP, p. 384). NATO Science and Technology Organization. https://www.sto.nato.int/publications/STO%20Technical%20Reports/Forms/Technical%20Report%20Document%20Set/docsethomepage.aspx?ID=4474&FolderCTID=0x0120D5200078F9E87043356C409A0D30823AFA16F6010066D541ED10A62C40B2AB0FEBE9841A61&List=92d5819c-e6ec-4241-aa4e-57bf918681b1&RootFolder=/publications/STO%20Technical%20Reports/STO-TR-SAS-114

NIC. (2007). *Iran: Nuclear Intentions and Capabilities* [National Intelligence Estimate]. National Intelligence Council. https://www.dni.gov/files/documents/Newsroom/Reports%20and%20Pubs/20071203_release.pdf

ODNI. (2022). *Intelligence Community Directive 203, Analytic Standards* (No. ICD-203). https://www.odni.gov/files/documents/ICD/ICD-203_TA_Analytic_Standards_21_Dec_2022.pdf

Ostermann, T., Vogel, H., & Appelbaum, S. (2018). Verbal Probabilities: Linear or Logistic? - A Regression Analysis Approach. *Studies in Health Technology and Informatics, 253*, 117–121.

Pashakhanlou, A. H. (2018). Intelligence and diplomacy in the security dilemma: gauging capabilities and intentions. *International Politics, 55*(5), 519–536. https://doi.org/10.1057/s41311-017-0119-8

Pherson, R. H., Donner, O., & Gnad, O. (2024). *Clear Thinking: Structured Analytic Techniques and Strategic Foresight Analysis for Decisionmakers*. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-48766-8

Pherson, R. H., & Heuer, R. J. (Eds.). (2019). *Structured analytic techniques for intelligence analysis* (Third edition). SAGE, CQ Press.

Piercey, M. D. (2009). Motivated reasoning and verbal vs. numerical probability assessment: Evidence from an accounting context. *Organizational Behavior and Human Decision Processes, 108*(2), 330–341. https://doi.org/10.1016/j.obhdp.2008.05.004

Putnam, R. D. (1988). Diplomacy and domestic politics: the logic of two-level games. *International Organization, 42*(3), 427–460. https://doi.org/10.1017/S0020818300027697

Rathbun, B. C. (2007). Uncertain about Uncertainty: Understanding the Multiple Meanings of a Crucial Concept in International Relations Theory. *International Studies Quarterly, 51*(3), 533–557. https://doi.org/10.1111/j.1468-2478.2007.00463.x

Reagan, R. T., Mosteller, F., & Youtz, C. (1989). Quantitative meanings of verbal probability expressions. *Journal of Applied Psychology, 74*(3), 433–442. https://doi.org/10.1037/0021-9010.74.3.433

Rovner, J. (2017). *Fixing the Facts: National Security and the Politics of Intelligence*. Cornell University Press. https://doi.org/10.7591/9780801463136

Rumsfeld, D. H. (2002). *Secretary of Defense Donald Rumsfeld and Air Force General Richard Myers, Chairman, Joint Chiefs of Staff, briefed February 12 at the Pentagon*. U.S. Department of Defense. https://usinfo.org/wf-archive/2002/020212/epf202.htm

Shying, M. (2013). Auditors interpretations of “in-isolation” verbal probability expressions: A cross-national study. *The Influence of Environmental and Cultural Factors in International Accounting. American Accounting Associations (AAA) Annual Meeting 2013*. https://www.sec.gov/rules/concept/s70400/shying1.htm

Somes, G. W. (1986). The Generalized Mantel-Haenszel Statistic. *The American Statistician, 40*(2), 106. https://doi.org/10.2307/2684866

Spiegel. (1984, March 18). »Dieser Dilettanten-Verein«. *Der Spiegel, 12*. https://www.spiegel.de/politik/dieser-dilettanten-verein-a-81aa81f9-0002-0001-0000-000013509312

Sutherland, H. J., Lockwood, G. A., Tritchler, D. L., Sem, F., Brooks, L., & Till, J. E. (1991). Communicating probabilistic information to cancer patients: Is there ‘noise’ on the line? *Social Science & Medicine, 32*(6), 725–731. https://doi.org/10.1016/0277-9536(91)90152-3

Tavana, M., Kennedy, D. T., & Mohebbi, B. (1997). An Applied Study Using the Analytic Hierarchy Process to Translate Common Verbal Phrases to Numerical Probabilities. *Journal of Behavioral Decision Making, 10*(2), 133–150. https://doi.org/10.1002/(SICI)1099-0771(199706)10:2<133::AID-BDM255>3.0.CO;2-5

Teigen, K. H. (2001). When Equal Chances = Good Chances: Verbal Probabilities and the Equiprobability Effect. *Organizational Behavior and Human Decision Processes, 85*(1), 77–108. https://doi.org/10.1006/obhd.2000.2933

Teigen, K. H. (2022). Dimensions of uncertainty communication: What is conveyed by verbal terms and numeric ranges. *Current Psychology*. https://doi.org/10.1007/s12144-022-03985-0

Teigen, K. H., & Brun, W. (1995). Yes, but it is uncertain: Direction and communicative intention of verbal probabilistic terms. *Acta Psychologica, 88*(3), 233–258. https://doi.org/10.1016/0001-6918(93)E0071-9

Teigen, K. H., & Brun, W. (2000). Ambiguous probabilities: when doesp=0.3 reflect a possibility, and when does it express a doubt? *Journal of Behavioral Decision Making, 13*(3), 345–362. https://doi.org/10.1002/1099-0771(200007/09)13:3<345::AID-BDM358>3.0.CO;2-U

Teixeira, C., & Fialho Silva, A. (2009). *The Interpretation of Verbal Probability Expressions Used in the IAS/IFRS: Some Portuguese Evidence*. http://dspace.uevora.pt/rdpc/handle/10174/6262

Tetlock, P. E. (2005). *Expert political judgment: how good is it? How can we know?* Princeton University Press.

Tetlock, P. E., & Mellers, B. A. (2011). Intelligent management of intelligence agencies: Beyond accountability ping-pong. *American Psychologist, 66*(6), 542–554. https://doi.org/10.1037/a0024285

Treverton, G. F. (2022). Connecting Intelligence and Policy. *Survival, 64*(1), 45–50. https://doi.org/10.1080/00396338.2022.2032957

US Congressional Select Committee on Intelligence. (2004). *Report on the U.S. Intelligence Community's Prewar Intelligence Assessments on Iraq*. United States Senate. https://irp.fas.org/congress/2004_rpt/ssci_iraq.pdf

Villejoubert, G., Almond, L., & Alison, L. (2009). Interpreting claims in offender profiles: the role of probability phrases, base-rates and perceived dangerousness. *Applied Cognitive Psychology, 23*(1), 36–54. https://doi.org/10.1002/acp.1438

Vogel, H., Appelbaum, S., Haller, H., & Ostermann, T. (2022). The Interpretation of Verbal Probabilities: A Systematic Literature Review and Meta-Analysis. In R. Röhrig, N. Grabe, V. S. Hoffmann, U. Hübner, J. König, U. Sax, B. Schreiweis, & M. Sedlmayr (Eds.), *Studies in Health Technology and Informatics*. IOS Press. https://doi.org/10.3233/SHTI220798

Wallsten, T. S., Budescu, D. V., Zwick, R., & Kemp, S. M. (1993). Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bulletin of the Psychonomic Society, 31*(2), 135–138. https://doi.org/10.3758/BF03334162

Wallsten, T. S., & Jang, Y. (2008). Predicting binary choices from probability phrase meanings. *Psychonomic Bulletin & Review, 15*(4), 772–779. https://doi.org/10.3758/PBR.15.4.772

Waltz, K. N. (1979). *Theory of international politics*. Addison-Wesley Pub. Co.

Wiener, J. L., & Mowen, J. (1986). SOURCE CREDIBILITY: ON THE INDEPENDENT EFFECTS OF TRUST AND EXPERTISE. *ACR North American Advances*. https://www.semanticscholar.org/paper/SOURCE-CREDIBILITY%3A-ON-THE-INDEPENDENT-EFFECTS-OF-Wiener-Mowen/9fa278e6f07d5e9dce15f2afc70df74f443ac3c8

Woodward, B. (2004). *Plan of attack*. Simon & Schuster.

Wyden, P. (1979). *Bay of Pigs: the untold story* (1st Touchstone ed). Simon and Schuster.

List of Appendices

Appendix no. 2: Scenario Selection Algorithm

The scenario selection algorithm determined the course of the survey, i.e. which tasks would be presented in what scenarios, depending on each participant's answers. The algorithm was implemented in the survey creation and hosting service *LimeSurvey* and uses the *PHP* scripting language.

Step	Description
1.1	A participant's expertise level on the subject area NATO is elicited (question code "ExpNATO"). The options range from 1 ("rudimentary awareness") to 5 ("recognised expert knowledge"). The expertise level selected by the participant is recorded as either "Exp1", "Exp2", "Exp3", "Exp4" or "Exp5" – depending on the level chosen.
1.2	The variable "ExpSum01" is calculated based on the expertise level chosen in question "ExpNATO": ExpSum01 = sum(2*countif("Exp2", ExpNATO), countif("Exp3", ExpNATO), 3*countif("Exp4", ExpNATO))
2.1-3.2	These two steps are repeated for each of the remaining two subject areas ("ExpUKR" and "ExpCHN"). The calculated sums are saved as "ExpSum02" for the expertise level on the Ukraine war and "ExpSum03" for the expertise on Taiwan/China.
4	The selection algorithm now determines which scenario a participant will be presented with in phase 3 of the survey based on the three calculated sums ("ExpSum01-03"). If a participant either reported the lowest or the very highest level of expertise for a subject area, this scenario is to be avoided. Among the remaining subject areas, the scenario is to be preferred, in which the participant's level of expertise is the 'most mediocre' – i.e. closest to level "Exp3".

If a participant answered either “Exp1” or “Exp5” for all subject areas, phase 3 of the survey is skipped altogether. Should a participant have a self-reported expertise level of “Exp3” in multiple subject areas, one of the corresponding scenarios is chosen randomly.

The algorithm’s code uses 26 nested if-functions:

```
SelAlgo = if(Exp == 0, if(min(ExpSum01, ExpSum02, ExpSum03) ==
ExpSum03, if(min(ExpSum01, ExpSum02, ExpSum03) == ExpSum02,
if(min(ExpSum01, ExpSum02, ExpSum03) == ExpSum01, rand(1, 3),
rand(2, 3)), if(min(ExpSum01, ExpSum02, ExpSum03) == ExpSum01,
sum((rand(0, 1) * 2), 1), 3)), if(min(ExpSum01, ExpSum02, ExpSum03) ==
ExpSum02, if(min(ExpSum01, ExpSum02, ExpSum03) == ExpSum01,
rand(1, 2), 2), 1)), if(Exp == 1, if(min(ExpSum01, ExpSum02) ==
ExpSum02, if(min(ExpSum01, ExpSum02) == ExpSum01, rand(1, 2), 2),
1), if(Exp == 10, if(min(ExpSum01, ExpSum03) == ExpSum03,
if(min(ExpSum01, ExpSum03) == ExpSum01, sum((rand(0, 1) * 2), 1), 3),
1), if(Exp == 11, 1, if(Exp == 100, if(min(ExpSum02, ExpSum03) ==
ExpSum03, if(min(ExpSum02, ExpSum03) == ExpSum02, rand(2, 3), 3),
2), if(Exp == 101, 2, if(Exp == 110, 3, if(Exp == 111, if(ExpNATO !=
"Exp5", if(ExpUKR != "Exp5", if(ExpCHN != "Exp5", rand(1, 3), rand(1, 2)),
if(ExpCHN != "Exp5", sum((rand(0, 1) * 2), 1), 1)), if(ExpUKR != "Exp5",
if(ExpCHN != "Exp5", rand(2, 3), 2), if(ExpCHN != "Exp5", 3, 4)))))))))))
```

The output of “SelAlgo” then determines which scenario is displayed in phase 3:

SelAlgo == 1 → Poland/NATO scenario in phase 3
SelAlgo == 2 → Crimea/Russia scenario in phase 3
SelAlgo == 3 → Taiwan/China scenario in phase 3
SelAlgo == 4 → Phase 3 is skipped, and the participant will directly proceed to the Taiwan/China scenario in phase 4

5 For all cases where SelAlgo < 4, the selection of the scenario for phase 4 will be determined by the selected scenario in phase 3:

SelAlgo== 1 (Poland/NATO in phase 3) → Taiwan/China in phase 4

SelAlgo== 2 (Crimea/Russia in phase 3) → Poland/NATO in phase 4

SelAlgo== 3 (Taiwan/China in phase 3) → Crimea/Russia in phase 4

Appendix no. 3: Survey Content

This is a complete rundown of the survey’s contents:

1.1 Expertise

Please rate your level of expertise in the following subject areas:

1.1.1 Transatlantic alliance (NATO)

□ 1 | Rudimentary awareness
□ 2 | Fundamental knowledge
□ 3 | Advanced knowledge
□ 4 | Excellent knowledge
□ 5 | Recognised expert knowledge

1.1.2 Russia's War in Ukraine

□ 1 | Rudimentary awareness
□ 2 | Fundamental knowledge
□ 3 | Advanced knowledge
□ 4 | Excellent knowledge
□ 5 | Recognised expert knowledge

1.1.3 East Asian Security (China/Taiwan)

□ 1 | Rudimentary awareness
□ 2 | Fundamental knowledge
□ 3 | Advanced knowledge
□ 4 | Excellent knowledge
□ 5 | Recognised expert knowledge

1.2 Prior experience with estimative judgements

Before we start, please indicate your prior professional experience with estimative judgements. Here, estimative judgements are defined as conclusions of analyses, made under uncertainty. You can find an example of an estimative judgement at the bottom of this page.

1.2.1 I am or have been the consumer of briefings or reports containing estimative judgements on matters of national security or security policy in a professional context (e.g. political analysis products, situational assessments or intelligence reports).

□ No, never
□ Yes, but rarely
□ Yes, sometimes
□ Yes, often

1.2.2 I am or have been the producer of briefings or reports containing estimative judgements on matters of national security or security policy in a professional context (e.g. political analysis products, situational assessments or intelligence reports).

□ No, never
□ Yes, but rarely
□ Yes, sometimes
□ Yes, often

2.1.1 1/3 Transatlantic alliance | Estimate creation

2.1.1.1 "POLAND leaves NATO within the next ten years." – Please assess the likelihood of that statement coming true.

□ Highly unlikely
□ Unlikely
□ Likely
□ Highly likely

2.1.1.2 "It is [answer of 2.1.1.1] that POLAND leaves NATO within the next ten years." – How high is your analytic confidence in that assessment? *(i.e. how confident are you in your judgement?)*

□ Low
□ Moderate
□ High

2.1.1.3 Please express your assessment ([2.1.1.1]) as a percentage.
0% would mean impossibility and 100% absolute certainty – however, these two extremes can't be selected here.

"The likelihood that POLAND leaves NATO within the next ten years lies at ____ percent."

2.1.1.4 Based on your analytic confidence ([2.1.1.2]), please add a margin of error to your percentage estimate. *The resulting interval can't go below 0% or above 100%.*

"The likelihood that POLAND leaves NATO within the next ten years lies at [2.1.1.3]% plus or minus ____ percentage points."

2.1.2 1/3 Transatlantic alliance | Estimate submission

Imagine you are an analyst and have to present your estimative judgement to a superior – in what format would you prefer to submit your assessment?

□ It is [2.1.1.1] that POLAND leaves NATO within the next ten years.

□ The likelihood that POLAND leaves NATO within the next ten years lies between [2.1.1.3 – 2.1.1.4]% and [2.1.1.3 + 2.1.1.4]%.

□ It is [2.1.1.1] that POLAND leaves NATO within the next ten years. The confidence in that assessment is [2.1.1.2].

2.2.1 2/3 Russia's war in Ukraine | Estimate creation

2.2.1.1 "CRIMEA remains under RUSSIAN control from 2024 to 2026." – Please assess the likelihood of that statement coming true.

□ Highly unlikely
□ Unlikely
□ Likely
□ Highly likely

2.2.1.2 "It is [2.2.1.1] that CRIMEA remains under RUSSIAN control from 2024 to 2026." – How high is your analytic confidence in that assessment?

□ Low
□ Moderate
□ High

2.2.1.3 Please express your assessment ([2.2.1.1]) as a percentage.
0% would mean impossibility and 100% absolute certainty – however, these two extremes can't be selected here.

"The likelihood that CRIMEA remains under RUSSIAN control from 2024 to 2026 lies at ____ percent."

2.2.1.4 Based on your analytic confidence ([2.2.1.2]), please add a margin of error to your percentage estimate. *The resulting interval can't go below 0% or above 100%.*

"The likelihood that CRIMEA remains under RUSSIAN control from 2024 to 2026 lies at [2.2.1.3]% plus or minus ____ percentage points."

2.2.2 2/3 Russia's war in Ukraine | Estimate submission

Imagine you are an analyst and have to present your estimative judgement to a superior – in what format would you prefer to submit your assessment?

□ It is [2.2.1.1] that CRIMEA remains under RUSSIAN control from 2024 to 2026.

□ The likelihood that CRIMEA remains under RUSSIAN control from 2024 to 2026 lies between [2.2.1.3 – 2.2.1.4]% and [2.2.1.3 + 2.2.1.4]%.

□ It is [2.2.1.1] that CRIMEA remains under RUSSIAN control from 2024 to 2026. The confidence in that assessment is [2.2.1.2].

2.3.1 3/3 East Asian Security | Estimate creation

2.3.1.1 " CHINA attacks TAIWAN by the middle of this century." – Please assess the likelihood of that statement coming true.

□ Highly unlikely
□ Unlikely
□ Likely
□ Highly likely

2.3.1.2 "It is [2.3.1.1] that CHINA attacks TAIWAN by the middle of this century." – How high is your analytic confidence in that assessment?

□ Low
□ Moderate
□ High

2.3.1.3 Please express your assessment ([2.3.1.1]) as a percentage.
0% would mean impossibility and 100% absolute certainty – however, these two extremes can't be selected here.

"The likelihood that CHINA attacks TAIWAN by the middle of this century lies at ____ percent."

2.3.1.4 Based on your analytic confidence ([2.3.1.2]), please add a margin of error to your percentage estimate. *The resulting interval can't go below 0% or above 100%.*

"The likelihood that CHINA attacks TAIWAN by the middle of this century lies at [2.3.1.3]% plus or minus ____ percentage points."

2.3.2 3/3 East Asian Security | Estimate submission

Imagine you are an analyst and have to present your estimative judgement to a superior – in what format would you prefer to submit your assessment?

□ It is [2.3.1.1] that CHINA attacks TAIWAN by the middle of this century.

□ The likelihood that CHINA attacks TAIWAN by the middle of this century lies between [2.3.1.3 – 2.3.1.4]% and [2.3.1.3 + 2.3.1.4]%.

□ It is [2.3.1.1] that CHINA attacks TAIWAN by the middle of this century. The confidence in that assessment is [2.3.1.2].

3. Short break

Please add some personal information about you. This is optional, i.e. not a mandatory part of the survey.

3.1 Age bracket

□ under 20

□ 20-24

□ 25-30

□ 20-29

□ 30-34

□ 35-39

□ 40-44

□ 45-49

□ 50-54

□ 55-59

□ 60-64

□ 65-69

□ 70-74

□ 75-79

□ 80 or older

3.2 Please select the country you are from.

Please select the country you are from. If you live and/or work in a country different to your nationality for at least 5 years, select the country you are living/working in.

[Dropdown menu with 247 country options]

3.3 Current Occupation

Student or Academic

→ Academic field [10 options + “other” w/ free text input]

Public Service

→ Ministry

→ Department [3 options + “other” w/ free text input]

→ Government agency

→ Department [4 options + “other” w/ free text input]

→ Armed forces

→ Other [free text input]

Political party or party-affiliated organisation

→ Elected representative

→ Party official

→ Staffer of an elected representative or party

→ Party-affiliated foundation

→ Other [free text input]

Private Sector

→ Consulting (e.g. security policy, international relations or defence)

→ Analysis (e.g. OSINT)

→ Defence industry

→ Other [free text input]

3.5 If you are interested in receiving an update on the results of this research project, you can leave your Email-address here.

[free text input]

4. Estimate evaluation I

4A **Now, you get to see the assessments that five other participants submitted when asked about the prospects of Poland leaving NATO within the next ten years. The participants were selected randomly.**

However: Two of the participants reported a higher level of expertise than you at the beginning of the survey (i.e. higher than "[answer of 1.1.1]"). The other three participants reported the same level of expertise as you or lower.

Please select the two submissions that, in your view, might be the ones coming from participants with a higher level of expertise.

□ $\text{VPE}_{\text{original}}$

□ $\text{NUM}_{\text{original}}$

□ $\text{VPE+C}_{\text{original}}$

□ $\text{VPE}_{\text{altered}}$

□ $\text{NUM}_{\text{altered}}$

*Options were presented in a random order; mock-assessments were generated based on the participant's original estimates.

4B **Now, you get to see the assessments that five other participants submitted when asked about the prospects of Crimea remaining under Russian control from 2024 to 2026. The participants were selected randomly.**

However: Two of the participants reported a higher level of expertise than you at the beginning of the survey (i.e. higher than "[answer of 1.1.2]"). The other three participants reported the same level of expertise as you or lower.

Please select the two submissions that, in your view, might be the ones coming from participants with a higher level of expertise.

□ $\text{VPE}_{\text{original}}$

□ $\text{NUM}_{\text{original}}$

□ $\text{VPE+C}_{\text{original}}$

□ $\text{VPE}_{\text{altered}}$

□ $\text{NUM}_{\text{altered}}$

*Options were presented in a random order; mock-assessments were generated based on the participant's original estimates.

4C **Now, you get to see the assessments that five other participants submitted when asked about the prospects of China attacking Taiwan by the middle of this century. The participants were selected randomly.**

However: Two of the participants reported a higher level of expertise than you at the beginning of the survey (i.e. higher than "[answer of 1.1.3]"). The other three participants reported the same level of expertise as you or lower.

Please select the two submissions that, in your view, might be the ones coming from participants with a higher level of expertise.

□ VPE$_{\text{original}}$

□ NUM$_{\text{original}}$

□ VPE+C$_{\text{original}}$

□ VPE$_{\text{altered}}$

□ NUM$_{\text{altered}}$

*Options were presented in a random order; mock-assessments were generated based on the participant's original estimates.

5. Estimate evaluation II

5.1A **Once again, you are provided with the assessments of five other participants. This time, however, imagine you are a decision-maker and the question whether Poland leaves NATO within the next ten years is vital for your strategic planning.**

Which of these assessments would you find most useful to base decisions on?

□ Analyst A: VPE$_{\text{original}}$

□ Analyst B: NUM$_{\text{original}}$

□ Analyst C: VPE+C$_{\text{original}}$

□ Analyst D: VPE$_{\text{altered}}$

□ Analyst E: NUM$_{\text{altered}}$

*Options were presented in a random order; mock-assessments were generated based on the participant's original estimates.

5.2A Why did you select the assessment of the analyst [answer of 5.1A]?

*multiple, non-exclusive choice

□ It aligns with my own assessment

□ It differs from my own assessment

□ It is precise

□ It leaves room for ambiguity / It is not overly-specific

□ It is easily comprehensible

□ It treats uncertainties transparently

□ It contains a high analytic confidence

□ Other [free text input]

5.1B Once again, you are provided with the assessments of five other participants. This time, however, imagine you are a decision-maker and the question whether Crimea remains under Russian control from 2024 to 2026 is vital for your strategic planning.

Which of these assessments would you find most useful to base decisions on?

*Options were presented in a random order; mock-assessments were generated based on the participant's original estimates.

□ Analyst A: $\text{VPE}_{\text{original}}$

□ Analyst B: $\text{NUM}_{\text{original}}$

□ Analyst C: $\text{VPE+C}_{\text{original}}$

□ Analyst D: $\text{VPE}_{\text{altered}}$

□ Analyst E: $\text{NUM}_{\text{altered}}$

5.2B Why did you select the assessment of analyst [answer of 5.1B]?

*multiple, non-exclusive choice

□ It aligns with my own assessment

□ It differs from my own assessment

□ It is precise

□ It leaves room for ambiguity / It is not overly-specific

□ It is easily comprehensible

□ It treats uncertainties transparently

□ It contains a high analytic confidence
□ Other [free text input]

5.1C Once again, you are provided with the assessments of five other participants. This time, however, imagine you are a decision-maker and the question whether China attacks Taiwan by the middle of this century is vital for your strategic planning.

Which of these assessments would you find most useful to base decisions on?

□ Analyst A: $\text{VPE}_{\text{original}}$
□ Analyst B: $\text{NUM}_{\text{original}}$
□ Analyst C: $\text{VPE+C}_{\text{original}}$
□ Analyst D: $\text{VPE}_{\text{altered}}$
□ Analyst E: $\text{NUM}_{\text{altered}}$

*Options were presented in a random order; mock-assessments were generated based on the participant's original estimates.

5.2C Why did you select the assessment of analyst [answer of 5.1C]?

*multiple, non-exclusive choice

□ It aligns with my own assessment
□ It differs from my own assessment

□ It is precise
□ It leaves room for ambiguity / It is not overly-specific

□ It is easily comprehensible
□ It treats uncertainties transparently
□ It contains a high analytic confidence
□ Other [free text input]

6. Real-world format usage

*only displayed if participant did not answer “No, never” to question 1.2.2

Final question: You indicated that, in a professional context, you [rarely/sometimes/often] produce (or produced) estimative judgements. In which format/s do (or did) you express these assessments?

□ Non-standardised verbal probability expressions
(e.g. "probably", "doubtful" etc.)

*multiple, non-exclusive choice

□ Standardised words of estimative probability
(e.g. according to NATO guidelines)

□ Standardised or non-standardised words of estimative probability
with a statement on analytic confidence (e.g. high/moderate/low)

□ Point estimates (e.g. "75% likelihood")

□ Numeric ranges (e.g. "likelihood between 70% and 80%")

□ Other [free text input]