RUNNING HEAD: SCENARIOS

Scenario Generation and Scenario Quality using the Cone of Plausibility

Mandeep K. Dhami[1*], Lars Wicke[2], and Dilek Önkal[3]

[1]Middlesex University, London, UK

[2]London School of Economics and Political Science, UK

[3]Northumbria University, UK

*Send correspondence to:
Mandeep K. Dhami
Department of Psychology
Middlesex University
The Burroughs, Hendon
London, UK
NW4 4BT
E-mail: m.dhami@mdx.ac.uk

Acknowledgements

This research was funded by HM Government. We would like to thank David Wasely for his research assistance.

Highlights

- Although there are a plethora of techniques that purport to facilitate the generation of good quality scenarios to support futures work, few have been subject to systemic inquiry.
- We empirically examined the number, type and quality of scenarios generated use of the Cone of Plausibility technique which is advocated across a range of organisations.
- We found that scenarios were significantly more likely to be highly coherent than highly complete, plausible, relevant/pertinent, or transparent. There were no significant differences in the quality of different types of scenario (i.e., baseline, best-case, worst-case). However, scenarios generated first (regardless of their type) were more likely to be highly relevant/pertinent and transparent than scenarios generated afterwards.
- These findings have practical implications for the use of the Cone of Plausibility and other qualitative scenario generation techniques.

Abstract

The intelligence analysis domain is a critical area for futures work. Indeed, intelligence analysts' judgments of security threats are based on considerations of how futures may unfold, and as such play a vital role in informing policy- and decision-making. In this domain, futures are typically considered using qualitative scenario generation techniques such as the cone of plausibility (CoP). We empirically examined the quality of scenarios generated using this technique on five criteria: completeness, context (otherwise known as 'relevance/pertinence'), plausibility, coherence, and order effects (i.e., 'transparency'). Participants were trained to use the CoP and then asked to generate scenarios that might follow within six months of the Turkish government banning Syrian refugees from entering the country. On average, participants generated three scenarios, and these could be characterized as baseline, best case, and worst case. All scenarios were significantly more likely to be of high quality on the 'coherence' criterion compared to the other criteria. Scenario quality was independent of scenario type. However, scenarios generated first were significantly more likely to be of high quality on the context and order effects criteria compared to those generated afterwards. We discuss the implications of these findings for the use of the CoP as well as other qualitative scenario generation techniques in futures studies.

Keywords: Scenario generation, cone of plausibility, best and worst case, wildcards, intelligence analysis, forecasting, futures and foresight

Introduction

According to Börjeson, Höjer, Dreborg, Ekvall and Finnveden (2006, p. 726), scenario thinking can help users to consider three different sorts of questions about the future: "what will happen? What can happen? And how can a specific target be reached?" These are precisely the sorts of questions pondered by the intelligence community in its strategic and tactical security missions. Indeed, the intelligence analysis domain is a critical area for futures work, and the task of scenario generation is considered to be pivotal to the work of intelligence analysts (Schmertzing, 2021). Using scenarios, analysts form judgments of how futures may unfold and these inform policy- and decision-making for tackling security threats.

In recent years, however, the effectiveness of the intelligence community's ability to generate accurate scenarios has been criticized. For example, many Western Governments did not anticipate the Arab Spring that followed events in Tunisia during December 2010, and which has had far reaching consequences (e.g., Goodwin, 2011; Morrell, 2015). Socio-political, economic, technological and health-related developments such as Brexit, the US-China trade war, Russia's cyber influence operations and the global pandemic are keeping the global intelligence apparatus busy (Office of the Director of National Intelligence, 2021).

Although there are both quantitative and qualitative methods for generating scenarios in futures work (for reviews see Amer, Daim, & Jetter, 2013; Börjeson, et al., 2006), the intelligence community has tended to favor the latter. Qualitative approaches rely primarily on the knowledge and expertise of individuals (or groups) rather than on historical statistical data, simulation and extrapolation which are used in quantitative approaches (Bradfield, Wright, Burt, Cairns, & Van Der Heijden, 2005; Ducot & Lubben, 1980; Huss & Honton, 1987). Indeed, the fact that futures studies in general and intelligence problems in particular are characterized by high levels of uncertainty, complexity, unpredictability and dynamism arguably necessitates the use of qualitative methods which appear to be more prevalent in such circumstances (Rounsevell & Metzger, 2010; Van Notten, Rotmans, van Asselt, & Rothman, 2003).

In its efforts to help analysts think critically and avoid bias and error when performing analytic tasks such as scenario generation (for a review see Dhami, Belton, & Careless, 2016), the intelligence community promotes the use of so-called 'structured analytic techniques'. In the present paper, we examine the quality of scenarios generated using a structured analytic technique that some intelligence analysts are trained and encouraged to use, namely, the Cone of Plausibility (CoP). Scenario quality is often used to judge the credibility of scenarios and consequently their usefulness in anticipating futures (Amer et al., 2006). Recent research has demonstrated that training in scenario generation can also improve forecasting accuracy (Chang, Chen, Mellers, & Tetlock, 2016; Mellers et al., 2014). Before we present the method and findings of the present study, we describe this technique and provide a brief review of the ways in which scenario quality can be assessed.

The Cone of Plausibility

As mentioned above, the intelligence community typically applies qualitative methods for scenario generation. In a review of the structured analytic techniques that analysts are trained to employ, Dhami et al. (2016) identified several techniques that have the primary function of helping analysts to generate scenarios. These differ in terms of the expertise, time and resources required to implement them. Scenario generation techniques used by the intelligence community include alternative futures analysis, the CoP, morphological analysis, multiple scenarios generation, outside-in-thinking, simple scenarios, and (individual/unstructured, group/virtual structured) brainstorming. The CoP in particular, is recommended by the Professional Head of Defence Intelligence in the UK as a "quick win for busy analysts" (UK Ministry of Defence, 2013).

It is believed that the CoP can help analysts to overcome hindsight bias, which the UK Ministry of Defence (2013) defines as "The tendency to underestimate how surprising past events were, which makes future shocks seem less plausible than they actually are" (p. 5). However, it is notable that the CoP is not included in the US Government's (2009) analytic tradecraft handbook. And so the question is: Does the quality of scenarios generated by the CoP justify its use as a scenario generation technique?

The CoP was originally designed by Taylor (1988, see also 1990) and later further developed for the intelligence domain by Heuer and Pherson (2014). This technique is used to generate a range of alternative scenarios. Analysts are first required to define the question to be answered, which is typically about how a situation will develop over a given timeframe. Then, analysts must identify key drivers of the issue in question and make assumptions about how these drivers are likely to develop over the chosen timeframe. Strategies for identifying drivers include STEEPLED (social, technological, economic, environmental, political, legal, ethical, and demographic). Next, analysts must generate a most probable 'baseline' future scenario by extrapolating the assumptions made about the drivers. This baseline typically represents what would occur if things were 'business-as-usual'. Analysts are also expected to generate one or more alternative scenarios that are based on changing one or more of the drivers and/or assumptions. One of these alternatives should be an 'opportunities' or 'best case' scenario, illustrating an outcome more positive than expected. Other alternatives may be worst case scenarios and/or low probability 'wildcard' scenarios (UK Ministry of Defence, 2013).

The CoP can be used by individual analysts as well as by groups. It can be employed relatively rapidly with minimal resources because it can be conducted in one session (and one iteration) and does not require gathering further information from sources beyond those already available at the start of the process, and nor does it require supplementation by the use of other techniques. However, the CoP does require some training. In addition, as Dhami et al. (2016) note, the CoP relies on the analyst's ability to identify key drivers across a range of domains. It also appears to narrow analysts' thinking to a few scenarios that he/she subjectively considers to be the most probable, best case, worst case, or wildcard (least probable).

Although some analysts such as those in the UK are trained and encouraged to use the CoP to generate scenarios (see e.g., UK Ministry of Defence, 2013), this technique, like others, as critics point out, has not been sufficiently empirically tested (e.g., Dhami, Mandel, Mellers, & Tetlock, 2015). In fact, some researchers have pointed to the lack of research generally on the issue of how to improve scenario generation (e.g., Isaksen, 1998; Wicke, Dhami, Onkal, & Belton, 2019). Given their value as inputs into futures, foresight, forecasting, planning and decision-making processes (e.g., Chang et al., 2016; Mellers et al., 2014; Onkal, Sayım, & Gönül, 2013; Wicke et al., 2019; Wright & Goodwin, 1999; 2009), generating scenarios that facilitate anticipation is a promising avenue for research (Andreescu Gheorghiu, Zulean, & Curaj, 2013; Börjeson et al., 2006; Postma & Liebl, 2005).

Assessing Scenario Quality

Most studies on scenario generation use the quantity of generated ideas as a measure of the effectiveness of a scenario generation technique. For instance, Isaksen's (1998) review of the brainstorming technique found that 32 out of 50 studies used quantity as an evaluation criterion. Only 15 studies used quality as a measure of effectiveness, and 10 studies used originality (see also Al-Samarraie & Hurmuzan, 2018). There may be a positive association between scenario quantity and quality. Paulus, Kohn and Arditti (2011) found that asking participants to focus on generating a large number of ideas led to better quality ideas (but see Wicke et al. [2019] who recently found no such relationship).

The quality of scenarios generated can be judged along different dimensions. In a review of 17 publications, Amer et al., (2013) identified the following criteria used to judge scenario quality, listed here in order of popularity: plausibility, consistency/coherence, relevance/pertinence, creativity/novelty, completeness, transparency and importance. In addition, some have judged scenarios based on the redundancy of the ideas generated (e.g., Furnham & Yazdanpanahi, 1995; Litchfield, Fan & Brown, 2011; see also Al-Samarraie & Hurmuzan, 2018). Nowack, Endrikat and Guenther (2011) suggest that scenarios should be credible (in terms of internal validity and reliability), transferable (in terms of external validity), and legitimate (from the perspective of the users).

It could be argued that the CoP, with its emphasis on generating best and worst case scenarios may sometimes yield scenarios that fall short on some of the aforementioned criteria such as plausibility. This is because such scenarios provide information about the 'tail' or extreme boundaries of what could occur. On the other hand, the CoP may yield scenarios that measure highly on criteria such as coherence and transparency because of its emphasis on identifying drivers and being explicit about assumptions that are made about these drivers. Finally, the use of 'structured' strategies such as STEEPLED to help identify drivers may mean that the CoP yields scenarios that are high on criteria such as relevance/pertinence. These issues will be explored in the present study.

The Present Study

The main aim of the present study was to assess the quality of scenarios generated using the CoP technique which is used as a basis for futures work by the intelligence community. Specifically, we aimed to examine (1) scenario quality on a range of criteria (i.e., completeness, plausibility, context otherwise known as 'relevance/pertinence', coherence, and order effects otherwise known as 'transparency); (2) the relationship between scenario quality and scenario type (e.g., baseline, best case, worst case); and (3) the relationship between scenario quality and scenario order (i.e., first v. afterwards).

Method

Participants

Participants were 35 final year Business undergraduate students who elected to take a module on forecasting at a Turkish University. Twenty-one (60.0%) of the sample were female. The mean age of the sample was 21.63 (*SD* = 1.22). The students had studied foundations of foresight in addition to quantitative and qualitative approaches to forecasting. Additionally, given the prominence of refugee issues in Turkey, where the study was conducted, and where there were already 3.5 million registered refugees at the time the study was conducted, along with the continuous media exposure to the issues, the participants were well-informed about the focal topic. Students have been commonly used in forecasting and foresight studies as reliable proxies for other decision-makers (e.g., Bilgic, Dhami, & Önkal, 2018; Goodwin, Gönül, & Önkal, 2019; Remus, 1986; Trottier & Gordon, 2016), including intelligence analysts (e.g., Wicke et al., 2019).

Cone of Plausibility Technique Training

Participants were trained to use the CoP technique described by Heuer and Pherson (2014). This comprises four steps as follows: (1) Define the question to be answered. (2) Identify key drivers of the issue and make assumptions about how they are likely to develop. (3) Generate a 'baseline' future scenario by extrapolating the assumptions made in Step 2. (4) In addition, generate one to three alternative scenarios based on changing one or more of the assumptions. Include at least one optimistic or 'opportunities' scenario. We also included training on developing worst case scenarios and wildcard scenarios because this is emphasized in the training of UK analysts (UK Ministry of Defence, 2013). The worked example that formed the training is provided in the Appendix.

Analytic Task

The analytic task on the topic of Syrian refugees was selected because it was considered pertinent to participants, and it was believed that they would have some general knowledge of the topic given its relevance and media-coverage in the country in which they lived. The topic is also representative of the futures work done by strategic analysts.

The task instructed participants as follows: “We want you to imagine that the Turkish government decides to ban all Syrian refugees from entering Turkey, effective immediately. Using the scenario generation technique you have been taught, we would like you to think about the scenarios (as many as you can) that might occur as a consequence of this policy decision within the next 6 months. Please use your own knowledge, and, if you want, you can also read some of the background information on the Syrian refugee crisis that follows.”

The information sheet summarized how the civil war in Syria and the refugee crises started, what each involved party is fighting for, how Syrian people are affected and how Turkey and other countries responded to the crises. This information was drawn from a variety of credible online sources (e.g., the BBC, World Bank).

Participants were instructed to write each scenario on a separate page and to number their scenarios in the order they generated them.

Procedure

Participation was voluntary, without any remuneration, and included informed consent and debriefing procedures. Data was collected individually. Participants were seated on separate tables in a typical auditorium, equipped with pencil and paper. Participants first received a tutorial by the third author on how to use the CoP technique, followed by a Q&A session. Participants all indicated that they understood the technique and had no further questions. Participants were then presented with the analytic task described above and had the opportunity to ask questions before starting. Participants were free to leave when they felt they had finished writing their scenarios (however many that was). The maximum duration for participation was approximately 110 minutes.

Data Coding

Before coding the quality of each scenario, we counted the number of scenarios generated by each participant and categorized their type (e.g., baseline, best case, etc).

Two trained coders, who were not involved in the research design or data collection, then independently coded scenario quality using a specially developed and tested coding scheme. Scenario quality was assessed using criteria taken from a review of the extant literature on scenario generation. Only criteria that could be reliably coded using at least a 3-category/point system (i.e., 1 = low, 2 = medium, 3 = high) were included in the final version of the scheme, and this was assessed by several iterations of testing coding reliability across the two coders. The final five scenario quality criteria were: completeness, plausibility, context (otherwise known as 'relevance/pertinence'), coherence, and order effects (otherwise known as 'transparency'). These criteria and their agreed coding are described below. In addition, the coding scheme included a brief 'definition' of each criteria (e.g., for coherence this was "no inconsistency/contradictions").

Completeness of a scenario refers to the detail of the drivers, outcomes, assumptions, and provision of relevant background information. This variable was coded as 'high' if the scenario contained a description of the drivers as well as the outcomes, showed some consideration of assumptions, and included relevant background information. Completeness was coded as 'medium' if the scenario contained a description of the drivers and outcomes, but no assumptions or background information. Completeness was coded as 'low' if the scenario only contained a description of outcomes or drivers and/or assumptions or background information.

Plausibility of a scenario refers to the link between the drivers and outcomes in the scenario, as well as assumptions. This variable was coded as ‘high’ if the scenario made explicit/clear links between specific drivers and specific outcomes, as well as links between specific drivers, outcomes and the assumptions underlying them. Plausibility was coded as ‘medium’ if the scenario made links between drivers and outcomes and assumptions that were not totally clear/explicit or specific; and ‘low’ if the scenario made no links between drivers and outcomes.

Scenario *context* refers to the current social, economic, legal, political context, as well as history. This variable was coded as ‘high’ if the scenario referred to three or more contexts (e.g., political, economic, security); ‘medium’ if the scenario referred to two contexts; and ‘low’ if the scenario referred to only one or no context.

Coherence of a scenario refers to the logical flow of the argument presented in the scenario. This variable was coded as ‘high’ if the ideas contained in the scenario were all consistent with each other and not contradictory; ‘medium’ if the scenario contained an argument with one inconsistency/contradiction; and ‘low’ if the scenario contained more than one inconsistency/contradiction.

Finally, the *order effects* variable refers to the identification of 2^{nd} or 3^{rd} order effects in the scenario. This variable was coded as ‘high’ if the scenario specified the potential effects of all the outcomes; ‘medium’ if the scenario specified the potential effects of only some of the outcomes; and ‘low’ if the scenario did not mention the potential effects of any outcomes.

The two trained coders worked separately. They each sent their final set of codes to the first author who noted any disagreements. There were few disagreements because, as mentioned earlier, the coding scheme was developed via a reliability testing procedure. Any disagreements were resolved via discussions between the two coders. In the rare instance when this was not resolved, the final decision was made by the first author who has the most experience in qualitative coding.

Analysis and Findings

The 35 participants generated a total of 116 scenarios. The average length of each scenario was 152 words (ranging from 44 to 441). Thirty-one participants (88.6%) generated three scenarios, two generated four scenarios, one participant generated five scenarios and one other generated 10 scenarios. On average, participants generated 3.31 scenarios (*SD* = 1.23).

All but two participants generated at least one baseline scenario, one best case scenario and one worst case scenario. The two exceptions had not generated a baseline scenario. In addition, only two participants used the term wildcard when labelling their scenarios. Finally, 40.0% of participants generated the best case scenario first, 37.1% generated the baseline scenario first, and the reminder (22.9%) generated the worst case scenario first.

Our analysis focuses on the first baseline, first best case and first worst case scenario generated by each participant. Thus, we will analyze 103 out of the 116 scenarios generated.[1] Below, we describe the data analyses and findings in relation to the aims of the study.

Scenario Quality and Quality Criteria

Figure 1 shows the proportion of scenarios across participants that were of high, medium or low quality on each of the five quality criteria (i.e., completeness, plausibility, context, coherence, and order effects). Only a tiny minority of scenarios were judged to be low quality on any of the five criteria. The vast majority of scenarios were either high or medium quality on each of the five criteria. Specifically, 55.3% of scenarios scored high on

[1] Thus, we excluded the two wildcard scenarios generated by two participants, seven scenarios from the participant who generated 10 (i.e., one baseline, one best case and five worst case), two scenarios from the participant who generated five (i.e., one best case and one worst case), and two (worst case) scenarios generated by the two participants who did not generate a baseline scenario.

the completeness quality criterion because they contained a description of the drivers as well as the outcomes, showed some consideration of assumptions, and included relevant background information. Over half (58.3%) of the scenarios were scored medium on the context criterion, meaning that they referred to two contexts. A relatively equal proportion of scenarios scored medium (49.5%) and high (48.5%) on plausibility because they made links between drivers and outcomes and assumptions (with the high quality ones being specific and explicit and clear in such links). The majority (79.6%) of scenarios scored high on coherence because the ideas contained in them were all consistent with each other and not contradictory. Finally, 62.1% of scenarios scored medium in terms of identification of order effects because they specified the potential effects of only some of the outcomes.

FIGURE 1 ABOUT HERE

Pairwise McNemar tests were computed to examine if the scores on some quality criteria (i.e., completeness, context, plausibility, coherence, and order effects) were higher than on others. For these analyses, we excluded the few cases that were coded as low on each of the five criteria and we applied Bonferroni corrections to the alpha level. The analyses revealed that scenarios were statistically significantly more likely to be high quality on the coherence criterion compared to the other four criteria (for completeness: χ^2 [98] = 12.41, $p < .001$; plausibility: χ^2 [99] = 21.95, $p < .001$; context: χ^2 [99] = 32.09, $p < .001$; and order effects: χ^2 [100] = 34.24, $p < .001$). In addition, scenarios were significantly more likely to be high quality on the completeness criterion than either the context criterion ($N = 98$, $p = .002$) or the order effects criterion (χ^2 [99], $p = .001$).[2] There were no other statistically significant associations between the scenario quality and the quality criteria, *p*s > .005.

Scenario Quality and Scenario Type

[2] The binomial distribution was used for this comparison.

Figure 2 shows the proportion of each type of scenario (i.e., baseline, best case, worst case) that were of low, medium or high quality on each of the five quality criteria. Baseline scenarios were more likely to score high on the coherence criterion (81.8%) than the other criteria, and only about a third (30.3%) of baseline scenarios scored high on the context criterion. Similarly, worst case scenarios were also more likely to score high on the coherence criterion (82.9%) compared to the other criteria. Here, only about a third (31.4%) scored high on the order effects criterion. Finally, while best case scenarios were also more likely to score high on the coherence criterion (74.3%), these scenarios were also more likely to score high on the completeness criterion (60.0%). However, only about 42.9% of best case scenarios scored high on the order effects criterion. Thus, all three types of scenario tended to be coherent i.e., contain ideas that were all consistent with each other and not contradictory.

FIGURE 2 ABOUT HERE

Pairwise McNemar tests were computed to examine whether some *types* of scenario (i.e., baseline, best case, worst case) scored higher than others on each of the five quality criteria. We again excluded the few cases that were coded as low and applied Bonferroni corrections. These analyses revealed no statistically significant associations between scenario quality and scenario type on any of the five quality criteria, *p*s > .003.

Scenario Quality and Scenario Order

Pairwise McNemar tests (with low scores removed and Bonferroni corrections) revealed that best case scenarios were statistically significantly more likely to be generated first than worst case scenarios, χ^2 (70) = 5.83, p = .016. There was no significant difference in order of generation of either baseline and best case scenarios or of baseline and worst case scenarios, *p*s > .017.

Figure 3 shows the proportion of scenarios generated first versus the proportion generated afterwards that were of low, medium or high quality on each of the five quality criteria. Here, scenarios generated first or afterwards were both more likely to score high on the coherence criterion (88.6% and 75.0%, respectively) compared to the other criteria. In addition, first scenarios were more likely to score high (80.0%) on the completeness criterion.

FIGURE 3 ABOUT HERE

Pairwise McNemar tests (again with low scores removed and Bonferroni corrections applied) were computed to examine if the order in which scenarios were generated (i.e., first v. afterwards) was associated with scenario quality on each of the five quality criteria. These analyses revealed that scenarios generated first were statistically significantly more likely to be high quality on two of the five criteria used to measure quality than scenarios which came afterwards. Specifically, first scenarios were more likely to be high quality on the context criterion (χ^2 [101] = 9.14, p = .002) and the order effects criterion (χ^2 [101] = 11.36, p = .001) compared to scenarios that were generated afterwards. There was no statistically significant association between scenario order and scenario quality on the remaining three quality criteria, *p*s > .01.

Discussion

Futures work is central to the domain of intelligence analysis and analysts' judgments about future states of the world are used to inform security policy- and decision-making (US Government, 2009; UK Ministry of Defence, 2013; see also Schmertzing, 2021). In an effort to help analysts generate useful scenarios, intelligence organizations train them to use structured analytic techniques such as the CoP (UK Ministry of Defence, 2013). Indeed, a variety of organizations around the world encourage the use the CoP to understand and grapple with potential futures (e.g., American Alliance of Museums, 2020; College of Policing, 2020; Government of Alberta, Canada, 2020).

Despite its popularity, there has been little empirical research into the quality of the scenarios generated using the CoP technique. This is also surprising given that scenario quality is considered to be an indicator of scenario credibility (Amer et al., 2013) and training in scenario generation has been shown to improve forecasting accuracy (Chang et al., 2016; Mellers et al., 2014). We examined the quality of the scenarios generated using the CoP on a range of criteria, and measured the association between scenario quality and scenario type as well as scenario order. To our knowledge, the present study represents a rare attempt to systematically and empirically measure scenario quality in this way. Several findings emerged and before we discuss these we consider the methodology employed in the present study.

The present study employed a futures (analytic) task representative of one faced by strategic analysts. Although it could be argued that futures studies ought to focus on longer time horizons than the six months examined here, it is also the case that the time horizon partly depends on the issue being studied (see Brier, 2005). The security domain where the current study is located, is characterized by the interplay of a myriad of highly uncertain, complex and dynamic factors which can create futures that have far reaching consequences for humanity. Indeed, the time horizon from the lone street vendor setting himself on fire in Tunisia to protest his ill-treatment by local officials in December 2010 to the start of the civil war in Syria which led to the refugee crisis in Turkey that is the focus of the scenarios studied here, was only a few weeks, and has had (and continues to have) far reaching consequences. It is therefore not uncommon for futures researchers working in the security domain to develop scenarios focusing on both near-term and longer-term time horizons (e.g., Gaub, 2019, 2020).

In addition, as mentioned earlier, this task was of relevance and interest to the Turkish Business students in our sample who were undertaking a course on forecasting. Participants had minimal opportunity to practice using the CoP beyond the training they were given, and they provided responses in English (which is not their native language, but is the medium of communication at their university). Despite this, participants fairly easily and rapidly constructed at least three scenarios, which were mostly of medium or high quality.

Scenarios Generated Using the CoP

Here, we discuss the main findings and point to avenues for further research on the topic as well as provide practical suggestions for use of the CoP. The present study found that the CoP yielded scenarios that were typically rated as either high or medium quality on each of the five criteria used to measure quality, namely, completeness, plausibility, context (also known in the literature as relevance/pertinence), coherence, and order effects (transparency).

More specifically, we also found that scenarios generated using the CoP technique were significantly more likely to be of high quality on the coherence criterion compared to the other four criteria. Scenarios contained ideas which were all consistent with each other and not contradictory. After a review of the literature on scenario quality, Amer et al. (2013, p. 37) conclude that the internal consistency of scenarios is "the most important" scenario validation criterion. Improvements in forecasting skill observed by some researchers appear to have come partly from training individuals to produce coherent scenarios (Chang et al., 2016; Mellers et al., 2014). As Spaniol and Rowland (2015) point out, scenarios (such as those generated using the CoP) are plausible, future-oriented, narrative descriptions. Research in other domains suggests that people search for coherent narratives among the information available to them, even to the point of excluding (potentially useful) information that is inconsistent (Pennington & Hastie, 1993).

Nevertheless, the present findings also suggest that individuals using the CoP may need further guidance on constructing scenarios that are high quality on some of the other quality criteria. In particular, only around a third of scenarios were rated as high quality on the context (relevance/pertinence) and order effects (transparency) criteria. Although one may have expected that strategies such as STEEPLED to have aided the construction of scenarios that are high on the context criterion, it is also the case that a broad knowledge base is required to refer to three or more contexts. Similarly, although the CoP requires the identification of drivers and being explicit about assumptions, which should make it easier to construct scenarios representing a logical progression over time, mental effort is required to generate high quality scenarios that specify the potential 2nd and 3rd order effects of all the outcomes. Improving performance on the context and order effects criteria is arguably important because they can provide insights into potential developments across different areas/domains as well as how those developments may come about, which ultimately make scenarios more useful for anticipating futures (Durance & Godet, 2010).

The need for further guidance is underscored by the association we found between scenario order and scenario quality. Specifically, scenario order was significantly associated with the quality of scenarios on the aforementioned two quality criteria. First scenarios (regardless of their type) were more likely to be of high quality on the context (relevance/pertinence) and order effects (transparency) criteria, than later scenarios. From a practical perspective, performance on the context criteria may be improved by employing group-based CoP, where individuals all have different areas of knowledge and expertise. Performance on the order effects criteria may be improved by training individuals to use "backward inference" whereby past observations "are used as evidence to infer the process(es) that produced them" – learning to understand past cause-effect relations can be useful for causal thinking about the future (Einhorn & Hogarth, 1985, p. 312).

Finally, we found no significant differences in the quality of different types of scenario on any of the quality criteria. This is no doubt a welcome finding for proponents of the CoP because otherwise some types of scenario may be given greater weight, thus potentially biasing policy- and decision-makers. One possible explanation for this finding is that unlike some other popular scenario generation techniques such as brainstorming, the CoP requires scenarios to be anchored to an initial (baseline) scenario. However, only 13 (out of 35) participants in our study generated the baseline scenario first. Research ought to examine the effect of scenario type order to determine if anchoring on any type of scenario helps to ensure that all other types of scenarios are of equal quality. If so, greater emphasis could be placed on producing scenarios in a specific order, and/or encouraging this sort of anchoring in other scenario generation techniques.

The CoP can be applied by groups, and a group-based procedure may be particularly useful when a broad knowledge base is required. Although collaborative analysis is an increasingly common feature of intelligence analysis (Hackman, 2011), it has time and resource implications, and may not always lead to better performance than individual analysis (Dhami & Careless, 2015). Thus, research could examine if scenarios generated collaboratively using this technique are of better or worse quality than those generated individually.

The present findings also comment on other features of the CoP and its use. The CoP requires the generation of a baseline scenario first, followed by a best case scenario and a worst case or wildcard scenario. However, we found that the majority of participants did not generate the different types of scenarios in the order they were instructed to, potentially reflecting a tendency to start with an extreme scenario rather than a baseline case. It is likely that under less controlled settings, individuals using the CoP may deviate even further from its structured nature. Although this deviation did not appear to have an adverse impact on scenario quality in the present study, the importance of the CoP's structured approach should be further investigated so as to determine whether instructions should better emphasize the need to follow a specific structure or not.

Comments from participants in the present study after data collection suggested that they would have preferred to 'cut-straight-to-the-chase' in writing the scenarios. Some of them found the identification of drivers and assumptions time consuming. Research could further examine the extent to which the use of STEEPLED and identification of drivers and assumptions affects scenario quality. Indeed, some participants felt that the structured format of the CoP stifled their creativity. In a recent study, Wicke et al. (2019) found that good quality scenarios could be generated using a simple brainstorming technique. However, unless instructed to do so, individuals using such techniques do not necessarily generate different types of scenarios (e.g., best case, worst case) that some believe can be useful for comparative judgment, improved forecasting skill as well as justifiable decision-making (e.g., Chang et al., 2016).

We found that best case scenarios were as equally likely as baseline scenarios to be generated first. However, best case scenarios were significantly more likely to be generated first compared to worst case scenarios. This suggests that scenarios of positive valence may come to mind more freely than scenarios of negative valence. The CoP is partly differentiated from other scenario generation techniques by its emphasis on extrapolating from the baseline scenario and so research ought to examine to what extent starting with a best case scenario (rather than a baseline scenario as some participants did in the present study) can bias the scenarios that follow from these. As mentioned above, scenario order was associated with scenario quality on two of the quality criteria (i.e., context and order effects), and it is unclear how best case scenarios that offer more breadth of domain coverage and specification of potential 2nd and 3rd order effects of outcomes might rose-tint the lens through which subsequent scenarios are conceptualized.

The utility of generating best case and worst case scenarios is potentially questionable. Research suggests that this approach can be detrimental to forecasting and decision-making (Goodwin et al., 2019; but see Chang et al., 2016), and that optimistic scenarios may bias prediction relative to pessimistic scenarios (Newby-Clark et al., 2000). In addition, when coupled with a baseline scenario, as in CoP, this trinary situation may create a 'compromise effect', where the baseline is preferred over the two extremes (e.g., Simonson, 1989; Simonson & Tversky, 1992). Others have also raised concerns with the tendency to focus on the middle scenario among a set of three when doing futures work (e.g., Pillkahn, 2008). Nevertheless, many scenario generation techniques, including those used in the intelligence analysis domain, still encourage best-worst case thinking (see Heuer & Pherson, 2014).

Only two participants in the present study labelled any of their scenarios as wildcard. Both used these to describe negative events or outcomes. Although the CoP emphases the exploration of 'opportunities' or best cases, the generation of worst case or (negative) wildcard scenarios is arguably useful because it can aid in strategic planning to mitigate unwanted adverse events or outcomes. It is unclear why so few participants generated wildcard scenarios, although commentators recognize the inherent difficulty in thinking about improbable, high impact, and often negative events (see Hiltunen, 2006). Greater emphasis can perhaps be placed on providing guidance for generating wildcard scenarios. This can be done not only by considering the impact of failed assumptions but also by considering turning points in a trajectory. These turning points may be surprising (e.g., Mendonça, Pina e Cunha, Kaivo-oja, & Ruff, 2004; see also Hiltunen, 2006) and could provide valuable insights when doing futures work.

The different types of scenarios can vary on several dimensions such as probability, valence (positive or negative), and impact. Baseline scenarios may be of either positive or negative valence, and are likely to be of higher probability but lower impact than alternative (best case, worst case or wildcard) scenarios. This is because they often represent a projection forward of the current situation. By contrast, best case scenarios are of positive valence, and are likely to be of lower probability but higher impact than baseline scenarios. While worst case scenarios are of negative valence, they are also likely to be of lower probability and higher impact than baseline scenarios. Wildcard scenarios may be either positive or negative (e.g., Linz, 2012), although some assume they will be negative (e.g., UK Ministry of Defence, 2013), and will, like best case and worst case scenarios, be lower probability and higher impact than baseline scenarios. Although we did not directly measure the probability and impact dimensions of the scenarios, the present findings suggest that individuals using the CoP may find it fairly easy to rapidly construct scenarios varying in these dimensions. Research should directly test this assertion.

The present findings support the use of the CoP as far as intelligence analysis and other organizations wish to encourage the generation of high quality scenarios. However, further research is needed to examine the CoP and the scenarios generated using this technique. For instance, to what extent does the CoP reduce hindsight bias as assumed by the UK Ministry of Defence (2013)? Research has shown that other techniques used to debias analysts are ineffective (e.g., Dhami, Belton, & Mandel, 2019; Mandel, Karvetski, & Dhami, 2018). To what extent is the CoP useful as a forecasting tool? Recent research reports that scenarios generated using a simple brainstorming technique typically contained several forecasts even though participants were not instructed to make forecasts and, on average, 45% of these forecasts were later found to be accurate (Wicke et al., 2019; see also Chang et al., 2016; Mellers et al., 2014). How well do the scenarios generated using the CoP support decision-making? Indeed, as Wright, Bradfield and Cairns (2013) point out, there is a dearth of research examining the effect of scenarios generated using different techniques on decision-making.

Finally, the fact that the intelligence community has numerous scenario generation techniques to choose from and which vary in their time and resource implications (see Dhami et al., 2016), suggests that research could also be directed at comparing the performance of the CoP against other techniques. Indeed, this would serve to answer calls for an evidence-based approach to intelligence analysis (Dhami et al., 2015). The Office of the Director of National Intelligence (2021, p. 4) states “The complexity of the threats, their intersections, and the potential for cascading events in an increasingly interconnected and mobile world create new challenges for the IC.” Thus, the enormity of the task faced by the intelligence community makes clear the increasing imperative for it to adopt evidence-based methods for futures and foresight thinking.

References

Amer, M., Daim, T. U., & Jetter, A. (2013). A review of scenario planning. *Futures*, *46*, 23-40.

American Alliance of Museums (2020). *Centre for the future of museums blog*. https://www.aam-us.org/2010/09/01/future-studies-101-potential-futures/

Al-Samarraie, H., & Hurmuzan, S. (2018). A review of brainstorming techniques in higher education. *Thinking Skills and Creativity*, *27*, 78-91.

Andreescu, L., Gheorghiu, R., Zulean, M. & Curaj, A. (2013). Understanding normative foresight outcomes: Scenario development and the 'veil of ignorance' effect. *Technological Forecasting & Social Change, 80*, 711-722.

Bilgic, A., Dhami, M. K., & D. Önkal (2018). Towards a pedagogy for critical security studies: Politics of migration in the classroom. *International Studies Perspectives, 19,* 250-266.

Börjeson, L., Höjer, M., Dreborg, K. H., Ekvall, T., & Finnveden, G. (2006). Scenario types and techniques: towards a user's guide. *Futures*, *38*(7), 723-739.

Brier, D. (2005). Marking the future: A review of time horizons. *Futures, 37*(8), 833-848.

Bradfield, R., Wright, G., Burt, G., Cairns, G., & Van Der Heijden, K. (2005). The origins and evolution of scenario techniques in long range business planning. *Futures*, *37*(8), 795-812.

Chang, W., Chen, E., Mellers, B., & Tetlock, P. (2016). Developing expert political judgment: The impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judgment and Decision Making, 11*(5), 509-526.

College of Policing (2020). *Analytic techniques*. https://www.app.college.police.uk/app-content/intelligence-management/analysis/analytical-techniques/

Dhami, M. K., & Careless, K. E. (2015). Intelligence analysis: Does collaborative analysis outperform the individual analyst? *The Journal of Intelligence Analysis, 22*(3), 43-58.

Dhami, M. K., Belton, I. K., & Careless, K. (2016, August). Critical review of analytic techniques. In *Intelligence and Security Informatics Conference (EISIC), 2016 European (pp.*152-155) IEEE.

Dhami, M. K., Mandel, D. R., & Belton, I. (2019). The 'Analysis of Competing Hypotheses' in Intelligence Analysis. *Applied Cognitive Psychology,* 33, 1080-1090. DOI: 10.1002/acp.3550

Dhami, M. K., & Mandel, D. R. (2012). Crime as risk taking. *Psychology, Crime and Law, 18,* 389-403.

Dhami, M. K., Mandel, D. R., Mellers, B., & Tetlock, P. (2015). Improving intelligence analysis with decision science. *Perspectives on Psychological Science, 10*(6), 753-757.

Ducot, G., & Lubben, G. J. (1980). A typology for scenarios. *Futures*, *12*(1), 51-57.

Durance, P., & Godet, M. (2010). Scenario building: uses and abuses. *Technological Forecasting and Social Change*, *77*, 1488-1492.

Einhorn H. J., & Hogarth R. M. (1985). Prediction, diagnosis, and causal thinking in forecasting. In Wright G. (Ed.) *Behavioral decision making*. Boston, MA: Springer. https://doi.org/10.1007/978-1-4613-2391-4_14

Furnham, A., & Yazdanpanahi, T. (1995). Personality differences and group versus individual brainstorming. *Personality and Individual Differences*, *19*(1), 73-80.

Gaub, F. (2019). *Scanning the horizon: 12 scenarios for 2021*. Paris: European Union Institute Security Studies.

Gaub, F. (2020). *Conflicts to come. 15 scenarios for 2030*. Paris: European Union Institute Security Studies.

Goodwin, J. (2011). Why were we surprised (again) by the Arab spring? *Swiss Political Science Review, 17*(4), 452-456

Goodwin, P., Gönül, M. S., & Önkal, D. (2019). When providing optimistic and pessimistic scenarios can be detrimental to judgmental demand forecasts and production decisions. *European Journal of Operational Research, 273*, 992-1004.

Government of Alberta, Canada (2020). *Cone of plausibility*. https://colab.alberta.ca/Mtds/Look/Pages/Cone.aspx

Hackman, J. R. (2011). *Collaborative analysis: Using teams to solve hard problems.* San Francisco: Berrett-Koehler.

Heuer, R. J., & Pherson, R. H. (2014). *Structured analytic techniques for intelligence analysis*. Washington DC: CQ Press.

Hiltunen, E. (2006). Was it a wild card or just our blindness to gradual change? *Journal of Future Studies, 11*(2), 61-74.

Huss, W. R., & Honton, E. J. (1987). Scenario planning—what style should you use? *Long range planning*, *20*(4), 21-29.

Isaksen, S. G. (1998). *A review of brainstorming research: Six critical issues for inquiry*. Buffalo, NY: Creative Research Unit, Creative Problem Solving Group-Buffalo.

Linz, M. (2012). Scenarios for the aviation industry: A Delphi-based analysis for 2025. *Journal of Air Transport Management, 22,* 28-35.

Litchfield, R. C., Fan, J., & Brown, V. R. (2011). Directing idea generation using brainstorming with specific novelty goals. *Motivation and Emotion*, *35*(2), 135-143.

Mandel, D. R., Karvetski, C. W., & Dhami, M. K. (2018). Boosting intelligence analysts' judgment accuracy: What works, what fails? *Judgment and Decision Making,* 607-621.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Guracy, B., Fincher, K., … & Murray, T. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, *25*(5), 1106-1115.

Mendonça, S., Pina e Cunha, M., Kaivo-oja, J., & Ruff, F. (2004). Wild cards, weak signals and organisational improvisation. *Futures, 36*(2), 201-218.

Morrell, M. (2015). *The Great War of Our Time: The CIA's Fight against Terrorism--From al Qa'ida to ISIS*. New York: Twelve, Hachette Book Group.

Newby-Clark, I. R., Ross, M., Buehler, R., Koehler, D. J., & Griffin, D. (2000). People focus on optimistic scenarios and disregard pessimistic scenarios while predicting task completion times. *Journal of Experimental Psychology: Applied, 6*(3), 171-182.

Nowack, M., Endrikat, J., & Guenther, E. (2011). Review of Delphi-based scenario studies: Quality and design considerations. *Technological Forecasting & Social Change, 78,* 1603-1615.

Office of the Director of National Intelligence (2021). *Annual threat assessment of the US intelligence community*. https://www.dni.gov/files/ODNI/documents/assessments/ATA-2021-Unclassified-Report.pdf

Önkal, D., Sayım, K. Z., & Gönül, M. S. (2013). Scenarios as channels of forecast advice. *Technological Forecasting & Social Change, 80*(4), 772-788.

Paulus, P. B., Kohn, N. W., & Arditti, L. E. (2011). Effects of quantity and quality instructions on brainstorming. *The Journal of Creative Behavior*, *45*(1), 38-46.

Pennington, N., & Hastie, R. (1993). The story model for juror decision making. In Hastie, R. (Ed.), *Inside the juror: the psychology of juror decision making* (pp. 192-221). Cambridge: Cambridge University Press.

Pillkahn, U. (2008). *Using trends and scenarios as tools for strategy development: Shaping the future of your enterprise.* Wiley.

Postma, T. J. B. M., & Liebl, F. (2005). How to improve scenario analysis as a strategic management tool? *Technological Forecasting & Social Change, 72*, 161-173.

Remus, W. (1986). Graduate students as surrogates for managers in experiments on business decision making. *Journal of Business Research, 24*, 19-25.

Rounsevell, M. D. A., & Metzger, M. J. (2010). Developing qualitative scenario storylines for environmental change assessment. *WIRES Climate Change, 1,* 606-619.

Schmertzing, L. (2021). Gaps between foresight and policy: The US "Global Trends Report" series case study. *Futures, 127*. https://doi.org/10.1016/j.futures.2020.102676.

Simonson, I. (1989). Choice based on reasons: the case of attraction and compromise effects. *Journal of Consumer Research, 16*(2), 158-174.

Simonson, I., & Tversky, A. (1992). Choice in context: tradeoff contrast and extremeness aversion. *Journal of Marketing Research, 29*(3), 281-295.

Spaniol, M. J., & Rowland, N. J. (2019). Defining scenarios. *Futures and Foresight Science, 1*(1), 1-13.

Taylor, C. W. (1988). *Alternative world scenarios for strategic planning*. Carlisle Barracks, PA: Strategic Studies Institute U.S. Army War College. https://apps.dtic.mil/sti/pdfs/ADA201065.pdf

Taylor, C. W. (1990). *Creating strategic visions*. Carlisle Barracks, PA: Strategic Studies Institute, U.S. Army War College.

Trottier K., & Gordon, I. M. (2016). Students as surrogates for managers: Evidence from a replicated experiment. *Canadian Journal of Administrative Sciences, 35*, 146-161.

UK Ministry of Defence (2013). *Quick wins for busy analysts*. London, UK: Defence Intelligence, UK.

US Government (2009). *A tradecraft primer: Structured analytic techniques for improving intelligence analysis.* https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/Tradecraft%20Primer-apr09.pdf

Van Notten, P. W. F., Rotmans, J., van Asselt, M. B. A., & Rothman, D.S. (2003), An updated scenario typology. *Futures, 35*, 423-443.

Wicke, L., Dhami, M. K., Onkal, D., & Belton, I. (2019). Using scenarios to forecast outcomes of the Syrian Refugee Crisis. *International Journal of Forecasting.* https://doi.org/10.1016/j.ijforecast.2019.05.017

Wright, G., & Goodwin, P. (1999). Future-focussed thinking: combining scenario planning with decision analysis. *Journal of Multi-Criteria Decision Analysis*, *8*(6)*,* 311-321.

Wright, G. & Goodwin, P. (2009). Decision making and planning under low levels of predictability: enhancing the scenario method. *International Journal of Forecasting*. *25*(4), 813-825.

Wright, G., Bradfield, R., & Cairns, G. (2013). Does the intuitive logics method – and its recent enhancements – produce “effective” scenarios? *Technological Forecasting & Social Change, 80*(4), 631-642.

Figure 1. **Scenario quality by quality criteria (*n* = 103 scenarios)**

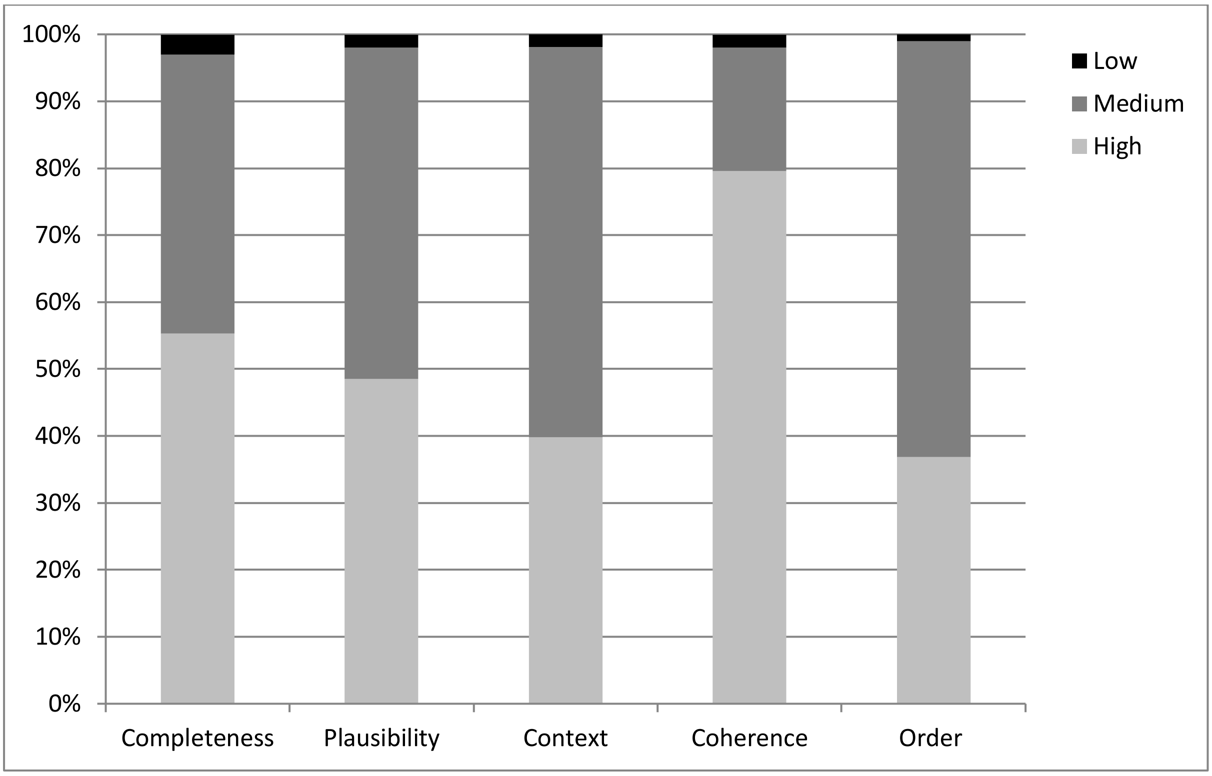


Figure 2. **Scenario quality by scenario type and quality criteria (*n* = 33 baseline scenarios, *n* = 35 best case and *n* = 35 worst case)**

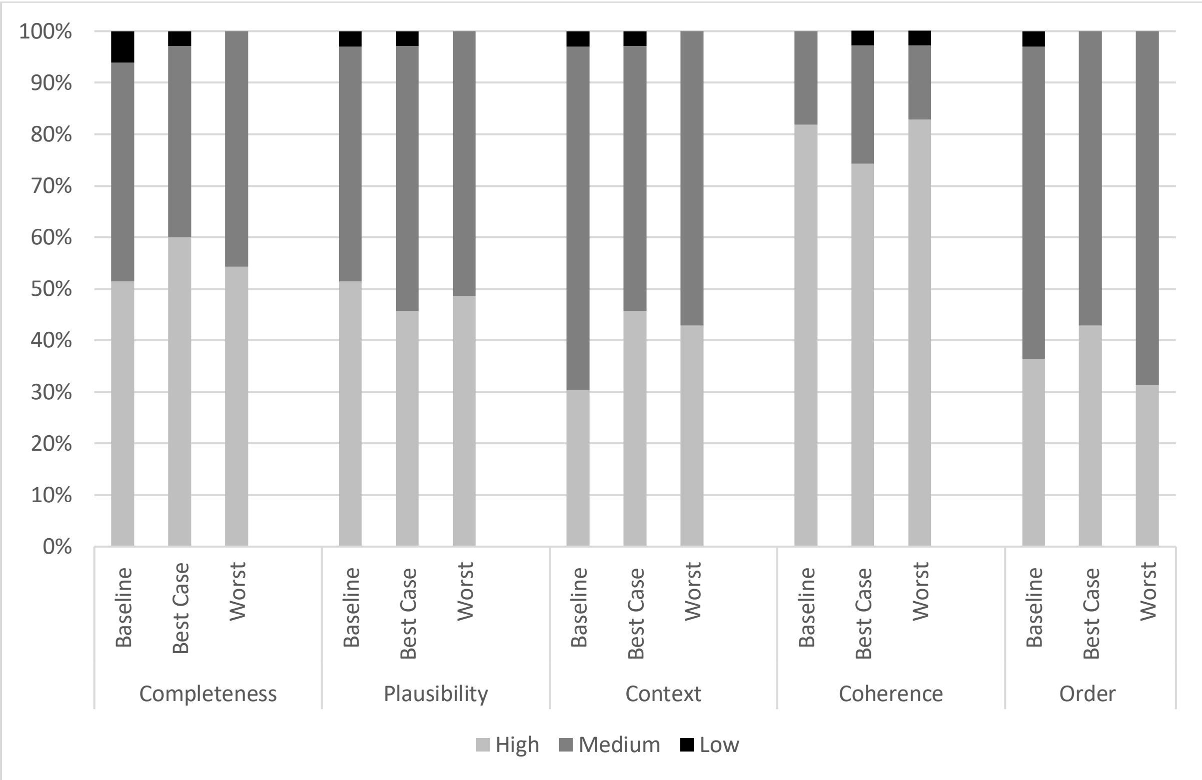
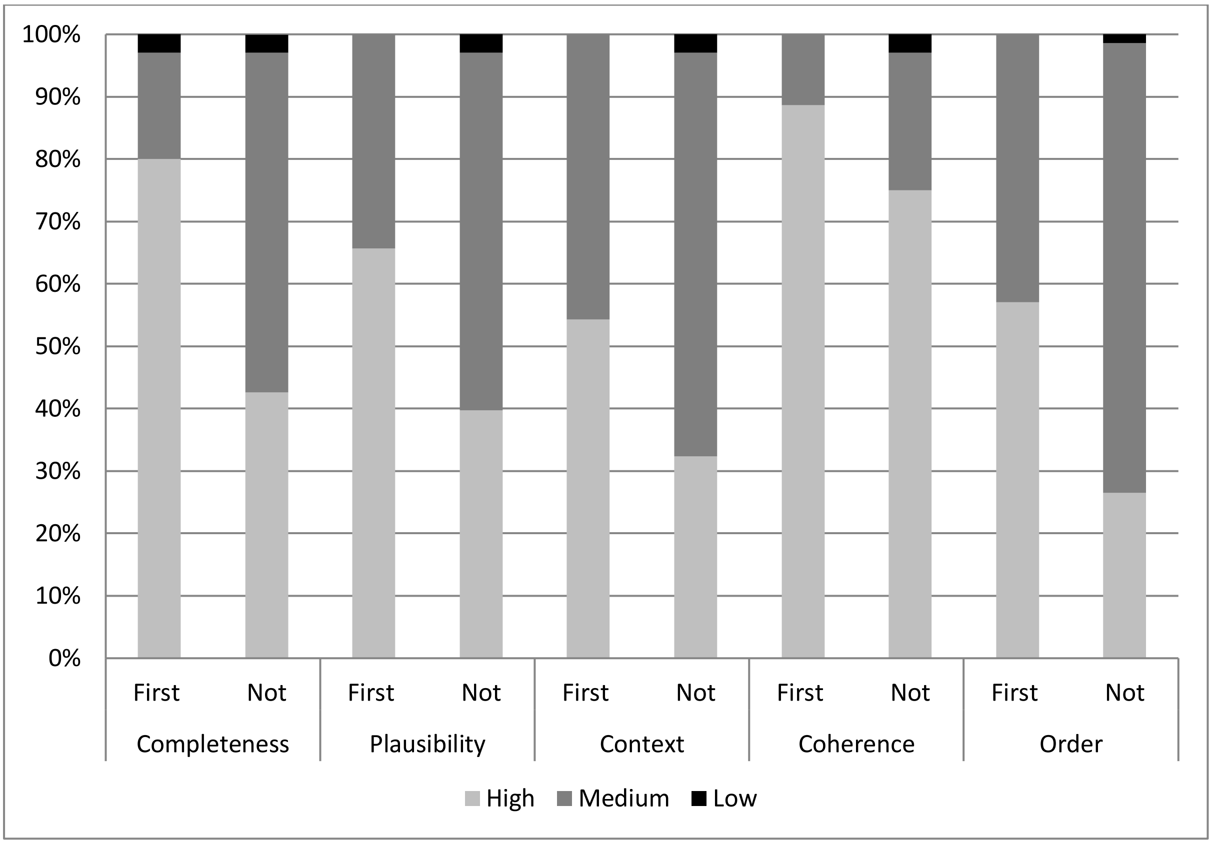


Figure 3. **Scenario quality by order and criteria (*n* = 103 scenarios)**



Appendix

CONE OF PLAUSIBILITY – TRAINING MATERIALS

Step 1: Define the question to be answered

"What are the implications for the UK by 2020 if it leaves the EU following a "leave" vote in the 2016 referendum?"

Step 2: Identify key drivers of the issue and make assumptions about how they are likely to develop.

The following key drivers were identified for the UK EU exit question:

- Trade with the EU
- Trade with other countries
- Immigration
- UK world influence
- UK sovereignty

Steps 3 and 4: Generate a "baseline" future scenario by extrapolating the assumptions made in Step 2, along with one to three alternative scenarios based on changing one or more of the assumptions. Include at least one optimistic "opportunities" scenario.