

The value of experiments in futures and foresight science: A reply

James Derbyshire¹, Mandeep K. Dhimi², Ian Belton³, Dilek Önkal⁴

1 Introduction

We thank those responding to our paper ‘The value of experiments in futures and foresight science as illustrated by the case of scenario planning’ (Derbyshire et al., 2022a). The research on which the paper was based was funded by a Research and Capacity Building Grant from the Society for the Advancement of Management Studies (SAMS) and the British Academy of Management (BAM). In addition to writing the focal paper, the project funded by the grant also saw the authors conduct three controlled experiments related to scenario planning, findings from which are to be reported in two further papers that are currently under review. We thank SAMS and BAM for their support in developing an area of research which is rapidly increasing in importance.

In the UK, scenarios featured centrally within official government decision-making processes during the coronavirus pandemic and were referred to in the media daily. The pandemic confirmed that, if it were ever possible to deny fundamental uncertainty or that new tools and methods are needed to tackle it, that is no longer the case. The dangers associated with closing the future to make it amenable to standard forecasting tools and probability-based decision aids are clear. Closing the future to make it amenable to these standard tools leads to blindsiding by surprises. Yet, recognising the dangers associated with tools that are presently more mainstream is one thing and evidencing the efficacy of alternative tools quite another. Experiments should be a tool in the methodological toolbox of those searching for this evidence for the reasons we outlined in our paper. Below we provide some reflections on four expert responses to our paper.

2 Valuing experiments in futures and foresight science: Reflections on four expert responses

Rowland and Spaniol (2022) highlight the potential for increased emphasis on experimentation to fragment and divide the field of FFS. But is this so if the underlying ontology adopted is one of critical realism rather than naïve empirical/logical positivism? There would indeed be potential for experiments underpinned by empirical/logical positivism to lead to divisions. Under the strict Popperian falsification-based view associated with empirical/logical positivism, a single experiment refuting a particular finding would falsify it. Yet, as reflected in the replication crisis facing the social and behavioural sciences, there are so few decisive tests, and as we noted, it is exceedingly difficult in an organisational setting to replicate exactly a previous experiment. Moreover, what really matters is the cumulative body of evidence and not the outcome of any one experiment. This is why we advocated the use of conceptual rather than direct replication—which, we believe, is less likely to lead to disagreements about the relative merits of others’ experiments that Rowland and Spaniol (2022) envisage.

From another perspective, we believe that divisiveness is useful if it facilitates discussion of disagreements and efforts to seek solutions. Much more problematic is a field in which everyone agrees. That said, we agree with Merton (1942) that scientific fields do not develop in the ‘rational’ way those adopting the empirical/logical positivist ontology think they do. As noted in Derbyshire (2017) by reference to Kuhn (1962), scientific fields tend to go through alternating periods of stability and disruption, which are reflected in periods of agreement and disagreement on theory and methods. For a new paradigm to emerge a field may need to undergo a period of disruption and disagreement.

¹ Chester Business School, University of Chester, corresponding author, email: j.derbyshire@chester.ac.uk

² Dept. of Psychology, Middlesex University

³ Strathclyde Business School, Strathclyde University

⁴ Newcastle Business School, Northumbria University

This leads us to ask several questions: In which part of this process of evolution is the field of scenario planning presently? And, similarly, what is the current evolutionary status of the field of FFS that subsumes scenario planning? Is there currently too much agreement rather than too little? Might a period of divisiveness not be welcome? If academia has a value that the domain of practice does not have, it is in its ability to generate paradigm shifts through disagreement, questioning and dialectics. Should these not be seen as a welcome by-product of increased emphasis on experimentation rather than as something undesirable? Perhaps greater divisiveness is just what is needed. Which brings us to Schoemaker's (2023) response to our paper.

We agree that Schoemaker's (1993) theoretical suggestion—i.e., that scenario planning's effect may be due to the conjunction fallacy—was not directly tested for in his 1993 paper. Yet, this theoretical suggestion emerged from the empirical tests that were conducted therein. We accepted the theoretical suggestion and have tested it in a fully controlled experiment conducted as part of our broader project, with the results to be outlined in a paper currently under peer review. Schoemaker's (2023) response is written as if our paper (Derbyshire et al., 2022a) is one that directly responded to and refuted Schoemaker (1993). Schoemaker (2023) gives the impression that our intention was simply to assert that the conjunction fallacy is not the basis for scenario planning's effect, without ever publishing the evidence underpinning that assertion. That was not the purpose of our paper. We should add that, it is the cumulative body of evidence that matters, not the findings from any one study (no matter how well it is conducted). We did not intend to suggest by briefly mentioning our forthcoming experimental findings that they are sufficient alone to falsify Schoemaker's (1993) theoretical suggestion about the conjunction fallacy. We were simply using our findings in relation to that suggestion to illustrate how experimental findings may provide incremental improvements to the 'weight of evidence' (Derbyshire et al., 2022b) and the direction in which it leans, which cumulatively (e.g., when added to those of Phadnis et al., 2015) determine what is currently thought to be true.

Schoemaker (2023) suggests that in Derbyshire et al. (2022a) we 'did not acknowledge the important follow-on experiments by Kuhn and Sniezek (1996) and falsely claimed that Phadnis et al. (2015) reported opposite results from Schoemaker (1993)' (Schoemaker, 2023, p.7). We did not acknowledge the important work of Kuhn and Sniezek (1996), of which we are very much aware, because, as we have stated, the purpose of the focal paper was not to write up our experimental findings. Kuhn and Sniezek (1996) are fully acknowledged in the paper that reports our experiment on the conjunction fallacy. More importantly, Schoemaker (2023), in suggesting here that we made a 'false claim', has made one of his own in relation to what we said about Phadnis et al. (2015). Firstly, we noted in our paper that Phadnis et al. (2015) did not use the same scenario process as Schoemaker (1993). Secondly, rather than claiming that Phadnis et al. (2015) 'reported opposite results', we stated that Phadnis et al. (2015) 'did not replicate' Schoemaker (1993).

This usefully highlights the difference between a conceptual replication—or 'conceptual extension' as Twang and Kwan (1999) alternatively call it (more on that subject later)—and a direct replication. When we said that Phadnis et al. (2015) 'did not replicate' Schoemaker (1993), we were referring to Phadnis et al. (2015) as a conceptual replication/extension of Schoemaker (1993), which it indeed was as acknowledged in Phadnis' (2022) response to our paper. Even if Phadnis et al. (2015) did not seek directly to replicate Schoemaker (1993), it was a conceptual replication/extension of it. If the conjunction fallacy really is the basis of an effect from scenario planning, we might expect a conceptual replication/extension such as that of Phadnis et al. (2015) to have shown similar results to those in Schoemaker (1993) that prompted this theoretical suggestion in the first place. This illustrates very nicely how conceptual replication/extension can add to the weight of evidence, which was the point we were trying to make by briefly referencing Schoemaker (1993) and Phadnis et al. (2015) in our paper. Whether we made that point clumsily or not is for others to judge, but we certainly did not make the 'false claim' suggested.

While our advocacy of conceptual replication may have lowered the bar in terms of exactly replicating what was done in an original study, it is still key to establishing a cumulative body of

evidence through experimentation that experiments be adequately reported. Their reporting should contain enough information for comparisons, caveats, and distinctions between studies to be drawn. This adequate reporting of experiments is needed because, as Schoemaker (2023, p.7) correctly notes, ‘claims about follow-on experiments either confirming or refuting early findings need to include appropriate caveats about ways in which the studies differed’. We agree wholeheartedly and therefore welcome the clarifications about Schoemaker (1993) in Schoemaker (2023). However, we suggest it might have been better to include these clarifications in Schoemaker (1993) in order to allow for the caveats to which Schomaker (2023) alludes. Claims about follow-on experiments either confirming or refuting early findings need to include appropriate caveats about ways in which the studies differed, exactly as Schoemaker (2023) states, but this can only happen if the original study (e.g., Schoemaker, 1993) provides the detail needed to make the caveats.

That said, we agree that many a managerial tool has waxed and waned down the years due to lack of empirical support. We agree that experiments are one way to defend a tool such as scenario planning against the criticism that it lacks empirical support, but we add two caveats in this regard. Firstly, the continued demand for scenario planning, which now spans decades, and the fact that it is a primary tool used by government for decision-making under uncertainty, suggests there is more to it than a mere managerial fad. There are very few tools for considering the future that do not require the future to be closed and scenario planning is one, which distinguishes it from others. Secondly, under the critical realist ontology we set out in Derbyshire et al. (2022a), to which many in FFS would adhere, experiments are just one method for evidencing the efficacy of a tool such as scenario planning. It is useful when used in conjunction with other tools such as case study, as we implied in our paper. We agree with Schoemaker (2023) about the need for thick descriptions through case studies, which should be carried out alongside experiments.

Moving onto the two remaining responses to our paper, Salo (2023) elaborates on and summarises very well the many difficulties associated with conducting experiments, and even more so replications of experiments, on a tool such as scenario planning that is used in highly idiosyncratic organisational contexts. An important danger highlighted by Salo (2023) is that of increased emphasis on experimentation giving rise to experiments that have limited correspondence with the settings in which scenario planning might naturally occur. In other words, there is a danger of conducting scenario-planning exercises under circumstances in which they would not otherwise occur, and which therefore badly replicate the real-world context in which they might happen, just for the sake of increased experimentation. Such artificially motivated scenario planning conducted simply to increase experimentation would lack external validity. It would certainly fall foul of the principles for designing field experiments highlighted by Phadnis (2022). It is for all these reasons that we emphasised the importance of representative design in our paper. See Dhami et al. (2004) for a broader discussion on the importance of representative design in relation to the experimental method.

The tensions at play here come starkly into focus when we consider that Salo (2023) quite rightly also highlights the level of granularity (and, therefore, the increasing sample size and effort) needed to test anything other than a quite general and high-level hypothesis, such as that related to the conjunction fallacy. Once one starts trying to take account of idiosyncratic contexts and their varying effect on the outcome from scenario planning, sample size and other considerations can become very taxing. In other words, once one attempts to isolate the specific procedures, contexts and circumstances associated with desirable effects by having a more elaborate experimental set up involving many dimensions of analysis, the effort needed to conduct experiments increases rapidly, and concomitantly, the feasibility of implementing it in a sufficient number of real settings decreases rapidly.

The tension here is between the need for external validity and the need for an experiment to be practically feasible. We highlighted this tension in relation to our discussion of online platforms as a tool for experimentation. On the one hand, they make experiments more feasible from a practical standpoint. On the other hand, however, experiments conducted using online tools are open to the accusation of lacking external validity when it comes to the testing of tools such as scenario planning

that are typically implemented offline, and which have a group-based social aspect that is difficult to replicate online. Rather than abandoning the experimental method, however, any issues around external validity and practical feasibility simply reinforce the need for a multi-pronged approach to research, which would include experiments conducted in both a real and online context, alongside case studies and action research, and other methods able to provide a deeper dive and assist in uncovering and fleshing out specific causal mechanisms, thereby generalising findings.

For still further reinforcement of the points made in Derbyshire et al. (2022a), we can draw on the work of Twang and Kwan (1999), who also consider replication and theory development from a critical realist perspective. In our paper we distinguished between direct and conceptual replication. Twang and Kwan (1999) make a still broader distinction between empiricism conducted for the various purposes of: checking of analysis, reanalysis of data, exact replication, conceptual extension (which we called conceptual replication), empirical generalisation, and generalisation and extension. These distinctions highlight the multi-faceted nature of empirical research, including that based on experiments. Twang and Kwan (1999) consider replication to be indispensable to theory development in management science. And as we did in our paper, Twang and Kwan (1999) also characterise knowledge development as a cumulative process in which it is the body of knowledge overall that develops. They recognise that no single study can stand alone as definitive.

Finally, the difficulties associated with conducting experiments on scenario planning again resurface in Phadnis (2022). Phadnis (2022) recognises the possibility for random assignment to treatment and control groups, which can increase what we referred to in our paper as internal validity, but also recognises that the random sampling needed to establish the external validity of an experiment is all but impossible in an organisational context. However, Phadnis (2022) proffers a potential solution to this problem. Namely, field experiments, which are experiments conducted in the natural setting in which the phenomenon under study occurs. Indeed, Phadnis et al. (2015) conducted a field experiment as well as a conceptual extension of Schoemaker (1993). Phadnis (2022) highlights the recognition in our paper of two threats to external validity—unrepresentative sample and unrepresentative design. The former is when a study's participants are not representative of the target population and the latter is when the experimental context is not representative of the real-world decision context. In response, Phadnis (2022) very usefully highlights Harrison and List's (2004) six factors for guiding the design of field experiments, which shaped the design in Phadnis et al. (2015). Field experiments that are cognisant of these six design factors can, we believe, increase the external validity of experiments in which random sampling is impossible. We again reiterate, however, that it is a multipronged approach to research, which would be achieved by those in the field using a range of empirical methods, that will ultimately lead to generalisable findings.

3 Conclusion

We believe the implication of our paper is a relatively uncontroversial and irrefutable one. Namely, that the FFS field—and the scenario planning subfield in particular—requires a step change in the variety of empirical studies conducted by those who research in it. We suggest that experiments are relatively neglected in this field in part because of the difficulty of conducting them in ever changing and idiosyncratic organisational contexts—a difficulty recognised in our paper and reinforced by the responses. Yet, if there is one thing in common across both the focal paper and its responses it is that the obstacles are *not* insurmountable. While it may well be impossible to conduct an experiment with full random assignment and random sampling, and one that fulfils all the other various requirements for representativeness and external validity discussed here and in the focal paper, it is nevertheless possible to conduct experiments that variously and partially meet some of these criteria, depending on the context in which they are conducted. To employ an apt cliché, we should not let the perfect get in the way of the good. In a field that is currently relatively lacking in experimental studies, even an experiment relatively lacking in many of these important regards can still contribute to advancing the field, can still garner many citations, and can still add to the weight of evidence.

And it is the cumulative weight of evidence that we wish to emphasise in conclusion. No one study or one researcher can expect or be expected to produce definitive findings that are forever set in stone. In that respect, we agree with the strict Popperian thesis in that all findings are provisional. The cumulative weight of evidence shifts over time and rests at any one time on many different studies, which employ a whole variety of methods of different types. We reiterate that we did not highlight experiments in our paper because we wish to place them on a scientific pedestal that, by definition, must diminish the usefulness and value of other empirical approaches. We highlighted experiments and made them the focus of our paper because we consider their paucity to be something that needs to be addressed by the field. To end on a positive note, this paucity is a golden opportunity for anyone who endeavours to rectify it.

References

- Derbyshire, J. (2017) Potential surprise theory as a theoretical foundation for scenario planning, *Technological Forecasting & Social Change* 82, 215-225.
- Derbyshire, J., Dhimi, M., Belton, I. & Önköl, D. (2022a) The value of experiments in futures and foresight science as illustrated by the case of scenario planning, *Futures and Foresight Science* (in press).
- Derbyshire, J., Feduzi, A. & Runde, J. (2022b) Borrowing from Keynes' *A Treatise on probability*: A non-probabilistic measure of uncertainty for scenario planning, *European Management Review* (in press).
- Dhimi, M. K., Hertwig, R. & Hoffrage, U. (2004) The role of representative design in an ecological approach to cognition, *Psychological Bulletin* 130, 959-988.
- Harrison, G & List, J. (2004) Field experiments, *Journal of Economic Literature* 42, 1009-1055.
- Kuhn, T. S. (1962) *The structure of scientific revolutions*, University of Chicago Press, Chicago.
- Kuhn, K. M. & Sniezek, J. A. (1996) Confidence and uncertainty in judgmental forecasting: Differential effects of scenario presentation, *Journal of Behavioral Decision Making* 9, 231-247.
- Merton, R. K. (1942). Science and technology in a democratic order, *Journal of Legal and Political Sociology* 1, 115–126.
- Phadnis, S. S. (2022) Considering field factors to enhance external validity of scenario planning experiments: A commentary on Derbyshire et al. 2023, *Futures & Foresight Science* (in press).
- Phadnis, S., Caplice, C., Sheffi, Y. & Singh, M. (2015) Effect of scenario planning on field experts' judgment of long-range investment decisions, *Strategic Management Journal* 36, 1401-1411.
- Popper, K. (1959). [2002] *The Logic of Scientific Discovery*, London: Hutchinson. Republished 2002. Routledge Classics.
- Rowland, N. & Spaniol, M. J. (2022) On replication in science: Commentary on Derbyshire et al. (2022), *Futures & Foresight Science* (in press).
- Salo, A. (2023) On the boundaries of experimental research on scenario planning, *Futures & Foresight Science* (in press).
- Schoemaker, P. J. H. (1993) Multiple scenario development: Its conceptual and behavioral foundation, *Strategic Management Journal* 14, 193-213.
- Schoemaker, P. (2023) Exploring scenario planning through controlled experimentation: Commentary on Derbyshire et al. 2023, *Futures & Foresight Science* (in press).

Tsang, E. W. K. & Kwan, K-M. (1999) Replication and theory development in organizational science: A critical realist perspective, *Academy of Management Review* 24, 759-780.