TRAFFIC
the wildlife trade monitoring network

Towards Robust Evaluations of Demand-Reduction Interventions:

Deep Dive Technical Briefing Paper on Social Science Surveys

Douglas MacFarlane[1]

This publication was produced with the financial support of the European Union. Its contents are the sole responsibility of TRAFFIC and do not necessarily reflect the views of the European Union

[1] PhD Candidate in Psychology | University of Western Australia; Visiting Researcher | Department of Zoology | University of Cambridge

1. INTRODUCTION

This Technical Briefing Paper aims to provide a 'deep-dive' analysis around social science surveys. The Paper explores positives, pitfalls and limitations of various approaches, with a focus on controlled and uncontrolled methods. To build on insight captured in other published literature focusing on experience and learning within the field of conservation, the paper presents case studies from the field of medicine.

2. IMPACT EVALUATION IN MEDICINE

2.1 Why medicine can teach us about demand reduction impact evaluation.

A significant amount of effort has been invested in determining the impact of different medical treatments on patients. This has resulted in better treatments but also abundant learning around the most effective approaches to social surveys in impact evaluation.

Consider the following examples;

1. A person suffering from osteoarthritis of the knee undergoes surgery, and a month later reports experiencing a significant reduction in pain.
2. A country that is a large consumer of shark-fin soup is exposed to a widespread public awareness raising campaign about the cruelty of this trade. Months later, consumers report consuming less shark fin soup.

In order to prove causality between the surgery/ campaign and reduction in pain/ consumption, adequate survey methods and safeguards need to be put in place. Whilst, most public health officials accept that valid conclusions regarding a treatment's effectiveness cannot be drawn in the absence of methodological safeguards i.e. against errors in causal inference (Lilienfeld, Ritschel, Lynn, Cautin, & Latzman, 2014); the same cannot be said in conservation (Sutherland et al., 2019: Veríssimo & Wan, 2019). Lessons from medicine can provide useful insight. However even in certain areas of modern medicine, this has not always been the case. Until relatively recently for instance, almost no

surgical procedures had ever gone through a rigorous clinical testing (Vedantam, Cohen, & Boyle, 2019).

One explanation for this, is that even seasoned surgeons and medical researchers can fall prey to psychological errors that make treatments seem effective, when they are not (Lilienfeld et al., 2014). The question about whether a particular surgery is effective can, understandably, seem sufficiently removed (and therefore appear not applicable) from interventions in the context of reducing demand for illegal wildlife products. However, when evaluating the impact of any intervention, the same principles about causal inference still apply. To demonstrate this, the next section will consider one of the most commonly performed orthopaedic surgeries.

2.2 Is keyhole surgery (arthroscopy) effective for treating arthritis of the knee?

Osteoarthritis is a common degenerative condition of the knee, causing the joint to become painful and stiff. To provide relief from these symptoms, orthopaedic surgeons will often recommend arthroscopic knee surgery. This minimally invasive "keyhole" surgery is expensive (more than 2 million annual procedures cost roughly USD $3bn in the US alone; (Siemieniuk et al., 2017)) and, as with all surgery, introduces some health risks. Given the large costs and potential harms, how is it possible to determine whether arthroscopic surgery is effective for treating osteoarthritis of the knee?

2.3 Uncontrolled consumer perception surveys

Participant perceptions:

One approach to evaluating an intervention is to conduct an uncontrolled survey asking direct questions. E.g. *Did your pain decrease after arthroscopic knee surgery?* When previous studies had asked this question, they found that roughly half of all patients who had arthroscopic surgery for osteoarthritis reported experiencing pain relief (Moseley et al., 2002). However—this question does not help determine whether this surgery is effective. The question instead measures people's perceptions of whether the surgery was effective.

Although the results may still be interesting, the evidence is limited to questions such as:

(i) Do patients believe that this surgery is responsible for relieving their pain?

(ii) Could some aspect of the surgery be providing pain relief, and

(iii) What experiences are patients likely to share with others? Thus, what narratives might exist to influence others who are deciding whether to have this surgery.

Expert perceptions: Another approach to impact evaluation is to conduct uncontrolled surveys of the perceived professional experts who advocate for the benefits of a particular intervention (for an example relevant to the wildlife trade, see Cheung, Mazerolle, Possingham, & Biggs, 2018). In the case of arthroscopy, the experts are the orthopaedic surgeons who diagnose the illness and, prescribe and perform the surgery on numerous patients. Gathering expert opinion would therefore involve surveying orthopaedic surgeons and asking: *Is arthroscopic knee surgery effective for treating osteoarthritis?* In a hypothetical example, it may be that all surgeons reported the surgery was effective for most of their patients—however, this does not provide evidence that the surgery is effective Just because surgeons are experts at performing the surgery and report that it can sometimes be effective, does not provide evidence that the surgery is effective. Psychological biases—such as confirmation bias, the illusion of control, and illusory causation (explained below)—conspire to deceive experts into believing that particular interventions are effective, when in fact, they are not (Lilienfeld et al., 2014). Conclusions are instead limited to:

(i) Do surgeons believe that the surgery can be effective?

(ii) What proportion of surgeons might recommend this surgery? and

(iii) Could the belief by orthopaedic surgeons in the perceived effectiveness of this surgery help explain why patients are choosing to have this surgery?

To explore further the limitations of these uncontrolled survey methods, it may be useful to consider a hypothetical example. For example, if we ran an uncontrolled survey with a sample of twenty patients (or twenty surgeons) who had the surgery (or performed the surgery) and found that fifteen patients reported experiencing subsequent pain relief, the results would paint a persuasive picture that the surgery was effective; 75% of patients appear to have benefitted. However, this conclusion is misleading—because the effect of the surgery is illusory (See Box 1).

Box 1. The illusion of causality

To understand why uncontrolled surveys do not provide evidence of effectiveness, we need to understand the phenomenon that enables us to mistakenly link outcomes with unrelated causes, namely the illusion of causality. This illusion occurs when an event (the outcome), by mere coincidence, occurs shortly after an action (the assumed cause). This sequence often leads us to believe that one event caused the other (Matute, Yarritu, & Vadillo, 2011). For example, after falling sick to a common cold, many people will consume vitamin C supplements, believing that they will help them fight a cold.

*The belief that vitamin C speeds up recovery is perpetuated because people will generally recover from a cold after a few days—**regardless of what actions they take** (Hemila & Chalker, 2013).The myth is also strengthened because people more readily recall instances when they took vitamin C and then recovered from a cold, compared to instances when they did nothing and recovered all the same.[2] Also related to this recall bias, the myth that vitamin C is effective is further strengthened by a reporting bias—the tendency to share positive treatment outcomes more than average outcomes—which then distorts the information available to others (de Barra, 2017; Ioannidis, 2017). In other words, when we take an action a confirmation bias is leading us to seek evidence that confirms that vitamin C is effective, rather than contrary evidence that it is not.*

To reliably test whether vitamin C actually provides a health benefit, assumptions should be tested using experimental design (Figure 2). Specifically, people need to know the number of people who (i) took the product and experienced a benefit, (ii) took the product and experienced no benefit, (iii) did not take the product but still experienced the benefit and (iv) did not take the product and experienced no benefit. Information from on all these four outcomes enables us to confidently conclude whether a benefit is caused by the product or whether there was some other causal factor (e.g., immune system).

Figure 2. A "contingency table" showing the summarised results of an experiment evaluating the effectiveness of vitamin C for treating common colds, compared to taking a sugar pill (i.e., a placebo). **The results show that 3 out of 4 people experienced the same health benefit after taking a vitamin C as after taking a sugar pill**.

	Outcome Present (e.g., Health benefit) 😀	**Outcome Absent** (e.g., No benefit) 😐
Cause Present (e.g., Taking a vitamin C pill)	3	1
Cause Absent (e.g., Taking a sugar pill believed to be vitamin C)	3	1

[2] Positive recall bias is linked to the illusion of control (Barberia, Blanco, Cubillas, & Matute, 2013). Specifically, when people take an action (taking vitamin C) they then actively anticipate and look for any possible effects (e.g., feeling better, side effects[2] etc.), which subsequently leading to greater recall. Whereas when people take no action, they do not look for possible effects, leading to lesser recall.

2.4 How can we test whether a surgery is effective? — Use an experimental design

Armed with a basic understanding of the illusion of causality, we can appreciate the need to run an experiment to compare the benefit of a health remedy with a placebo intervention (e.g., a sugar pill). But how can we apply a placebo control to evaluate a surgery?

To test whether arthroscopic knee surgery caused patients to experience pain relief, researchers needed to run a controlled experiment using a sample of people suffering from osteoarthritis. For the results of the experiment to be convincing, the researchers needed to remove (known as controlling for) key sources of potential influence (referred to as confounding factors) that could have biased the results and thus suggested a misleading answer. The following section outlines some of the key confounding factors (i.e., influences other than the surgery) that might have contributed to people's experience of pain relief. The following also outlines underlying logic to explain how experimenters can exclude the possibility that each confounding explanation inadvertently influenced their experimental results.

Problem #1: Placebo effects — The pre-existing beliefs of patients. Patients may have had pre-existing beliefs about the effectiveness of arthroscopic surgery that might have influenced their experience of pain relief. For example, people who believed that the surgery was effective, may have been more likely to experience, or report experiencing, a benefit from having the surgery.

Solution #1*:* Before agreeing to participate in the experiment, people were made aware that they would be assigned to one of three treatment groups. Participants were also made aware that they would not know which treatment they would receive, even after the experiment. To ensure that patients were truly unaware of which treatment they had received, several steps were taken to ensure all participants shared a similar experience.

The crucial differences in the treatments were:

1. Treatment Group #1 (The real surgery): Patients were put to sleep, given incisions, had their knees flushed with saline, and then received the real arthroscopic knee surgery.
2. Treatment Group #2 (Saline wash only): Patients were put to sleep, given incisions, and then had their knees flushed with saline but did not have the surgery. This treatment group allowed the experimenters to test whether the saline wash alone could have been responsible for the experience of pain relief.
3. Treatment Group #3 (Control group or "placebo surgery"): Patients were put to sleep and were given incisions but did not receive the saline wash or the surgery.

NB: When participants in all three conditions woke up after the procedure, they all had the same scarring from the incisions and thus no means of knowing which treatment they received.

Box 2. Placebo vs Nocebo

*Much research demonstrates that when people believe they are being treated, even when they are given an inert or irrelevant treatment, the human body can respond in several important psychological ways that can produce a real impact on recovery and the experience of pain relief (Stewart-Williams, 2004). This phenomenon is called the **placebo** effect. Conversely, the opposite is also true, when people receive an inert or irrelevant treatment that they believe can cause health harm, then they are more likely to report and/or experience symptoms consistent with those perceived harms. This phenomenon is called the **nocebo** effect (Crichton & Petrie, 2015).*

Problem #2: Demand Characteristics — The pre-existing beliefs of hospital staff

The hospital staff delivering the surgery (e.g., the surgeon and nurses etc) may also have had pre-conceptions about the effectiveness of the surgery, which might have influenced how they treated some patients. The difference in treatment, attitudes, and interactions might then have altered some patients experience of pain relief.

Furthermore, if the hospital staff knew which treatment each patient received, then they may have accidentally, or even subconsciously, revealed this to the patients (e.g., via body language). This may not only have influenced patients' experience of the surgery but also how they reported their experiences. For example, if the patient knew they had the arthroscopic surgery, and then felt that the surgeon was eager to demonstrate that the surgery was effective, then patients might have been more inclined to report experiencing a benefit to avoid disappointing the surgeon.

***Solution #2*:** The surgeons and the nurses caring for the patient were "blinded" to which patients received which treatment before, during, and after the surgery. To achieve this in practice, the surgeon did not know who would receive what treatment. Instead, the surgeon was given an envelope immediately before the surgery directing them to perform one of the three treatments.

Problem #3: Selection Bias — Natural differences between the treatment groups.

Patients in the control group might have been different in some meaningful way to the patients in one or both of the treatment groups. Differences between treatment groups often arise when the selection process is somehow biased. For example, if the process tended to select younger patients for Treatment Group #1 than those selected for Treatment Groups #2 and #3, then the difference in age between the groups could have explained why Treatment Group #1 experienced a benefit where the other groups did not. In other words, if the selection process was biased, then the researchers might wrongly demonstrate that the surgery caused a benefit, when in fact, what caused the differences in pain relief was actually due to the age differences between the different groups.

***Solution #3*:** To exclude the possibility that individual factors (e.g., age, gender, height, weight, fitness, or existing beliefs about the surgery etc) might influence the outcome of the experiment—the process that decided which patients received what treatment needed to be random (e.g., a roll of the dice). By introducing randomisation into the selection process, this helps to ensure that individual-level differences are evenly distributed between the control and the treatment groups.

Box 3. Demand characterises in consumer surveys

In research, particularly psychology, a demand characteristic is a property of an experiment whereby participant's interpretation of the purpose of an experiment, alters their behaviour to fit that interpretation. However, demand characteristics can also undermine formative research (e.g., qualitative attitude surveys), which is particular challenge for research on consumers of wildlife products. Not least because many wildlife products are illegal and/or have become somewhat socially unacceptable, meaning that participants may alter their responses to avoid legal or social consequences (real or perceived).

For example, a recent survey of shark-fin soup consumers asked, 'Were you aware of these facts? Every year millions of sharks are butchered solely for their fins. In many instances the fins are cut off while the shark is alive, it is then thrown back into the sea to drown and die. How likely will you be to refuse eating shark fin soup?'.

The framing for this question signals to the survey participant that the researcher believes the shark fin trade is cruel. Participant responses may therefore have been influenced by the implicit purpose of the survey (i.e., to stop a cruel trade), meaning that the results may not be a true reflection of participant's beliefs, but instead their responses may have been influenced by the desire to please (or not upset) the researcher (see also preference falsification, (Sunstein, 2019)).

For such formative research, a better approach would be to blind the purpose of the survey to participants, by providing only neutral statements; i.e., a balanced, but random, mix of pro-shark fin and anti-shark fin statements and then asking participants to state their level of agreement with each statement. This way, participants would be unable to guess the purpose of the research thus helping to reduce this form of response bias. Unfortunately, this approach does not completely solve the problem of preference falsification. Instead, ideally would also seek to verify participant responses by comparing them with observations of real consumer behaviour (e.g., sales figures — see forthcoming Burgess & Broad, 2020) .

To check that the randomisation process has worked, such factors (e.g., age, gender) need to also be measured and compared, because even by random chance, one group may have turned out to be significantly different to the other, which could then unfairly bias the results.

2.5 Dispelling the illusion of causality — Experimental results

This understanding of the experimental design helps illustrate why the results of our first research question —"did you experience pain relief after the surgery?"—contributed an illusory effect of the surgery. If we asked the same questions using an experimental design, using a sample of 40 patients who believed they had the surgery, but only 20 having the real surgery and the other 20 having the "placebo" surgery and then reported the results as in Figure 3. When a fair comparison is made between the treatment and a comparable placebo treatment, people's experience of pain relief was the same whether or not they had the surgery—ie. the arthroscopic surgery was not the cause.

	Outcome Present (e.g., Health benefit) 😀	**Outcome Absent** (e.g., No benefit) 😐
Real surgery	15	5
Placebo-surgery	15	5

Figure 3. A visual summary of the hypothetical results of a controlled survey study. The results help to dispel the illusion of causality by demonstrating that 3 out of 4 people experienced the same benefit after having the placebo surgery as after having the real surgery.

The real results of the experiment on arthroscopic surgery for osteoarthritis revealed a total of 165 patients who completed the trial and were assessed at multiple points over 24 months using a self-reported pain scale. At no point in the experiment, did either treatment groups (arthroscopy or saline wash) report less pain or better function (walking and climbing stairs) than the control group (placebo-surgery). The experimenters concluded that neither arthroscopic surgery nor the saline wash provided any benefit over placebo surgery (Moseley et al., 2002).

2.6 What we can and cannot conclude from this experiment

Importantly, this experiment does not provide evidence that arthroscopic surgery is ineffective for other medical conditions. Furthermore, the experiment leaves open the possibility that there may be something about the process such as visiting a hospital, having incisions, and/or having a general aesthetic that may be causing people to experience pain relief and/or otherwise benefit from a natural placebo response.

Such results are typically what health professionals are referring to when they conclude that this surgery is not supported by the evidence. For accuracy, however, it would be better to state that—a randomised controlled trial found that this intervention had no benefit beyond a placebo effect. Importantly, if we had simply asked patients and surgeons whether the intervention was effective—our results would have created a convincing illusion of causality that the surgery had been effective. Furthermore, we would not have learned that any benefits derived from a placebo effect could also been achieved without the surgery.

2.7 Additional Caveat — Meaningful causal effects should be repeatable

A final point to note is that just because one experiment found no beneficial effect of the surgery this may not be conclusive evidence that the intervention is ineffective. Occasionally, by chance, even the best-designed experiments can produce two types of misleading results.

In the context of the arthroscopic surgery, false results are called either a (i) "false positive"—results that show a positive benefit of an ineffective surgery or (ii) a "false-negative"—results that show no benefit of an effective surgery (Reinhart, 2015). To increase our confidence in our conclusion that this particular surgery is not effective (i.e., avoid a false-negative), we need to re-run the experiment to test if we get similar results (i.e., test if the results are "replicated").

In 2015, researchers published a systematic review of nine experimental trials of arthroscopic knee surgery for osteoarthritis. The review found that arthroscopic knee surgery had no benefit on function and only a small but inconsequential effect on pain relief that was no longer present after one to two years.

Furthermore, they also found that nine studies reported on significant harms such as deep vein thrombosis, pulmonary embolism, infection, and even death. The researchers concluded that given the harms associated with surgery—arthroscopic surgery should no longer be used to treat osteoarthritis of the knee (Thorlund, Juhl, Roos, & Lohmander, 2015).

Thankfully, the necessity for experimental trials has been increasingly recognised by orthopaedic surgeons. Indeed, another systematic review published in 2015 assessed the effectiveness of invasive surgical procedures when tested using placebo-controlled experiments. Their meta-analysis showed that whilst invasive surgery can have a large impact on several health outcomes—for pain-related conditions (e.g., back pain and arthritis) they found no benefit of surgery compared to placebo surgery.

The researchers concluded that more experimental evidence was needed to 'avoid a continuation of ineffective treatments' (Jonas et al., 2015).

2.8 Conclusion: Research that conflates perceived causality with real impact is misleading

The conflation *between perceptions (i.e., of patients and surgeons) of what caused* some patients to experience pain relief and *what actually caused* some patients to experience pain relief; is similar to the confusion between what interventions *are effective in changing behaviour*, and what interventions *people perceive to be effective in changing behaviour*".

To summarise in the major points outlined thus far, Figure 4 provides a visual representation of some key differences between uncontrolled surveys and an experimental approach.

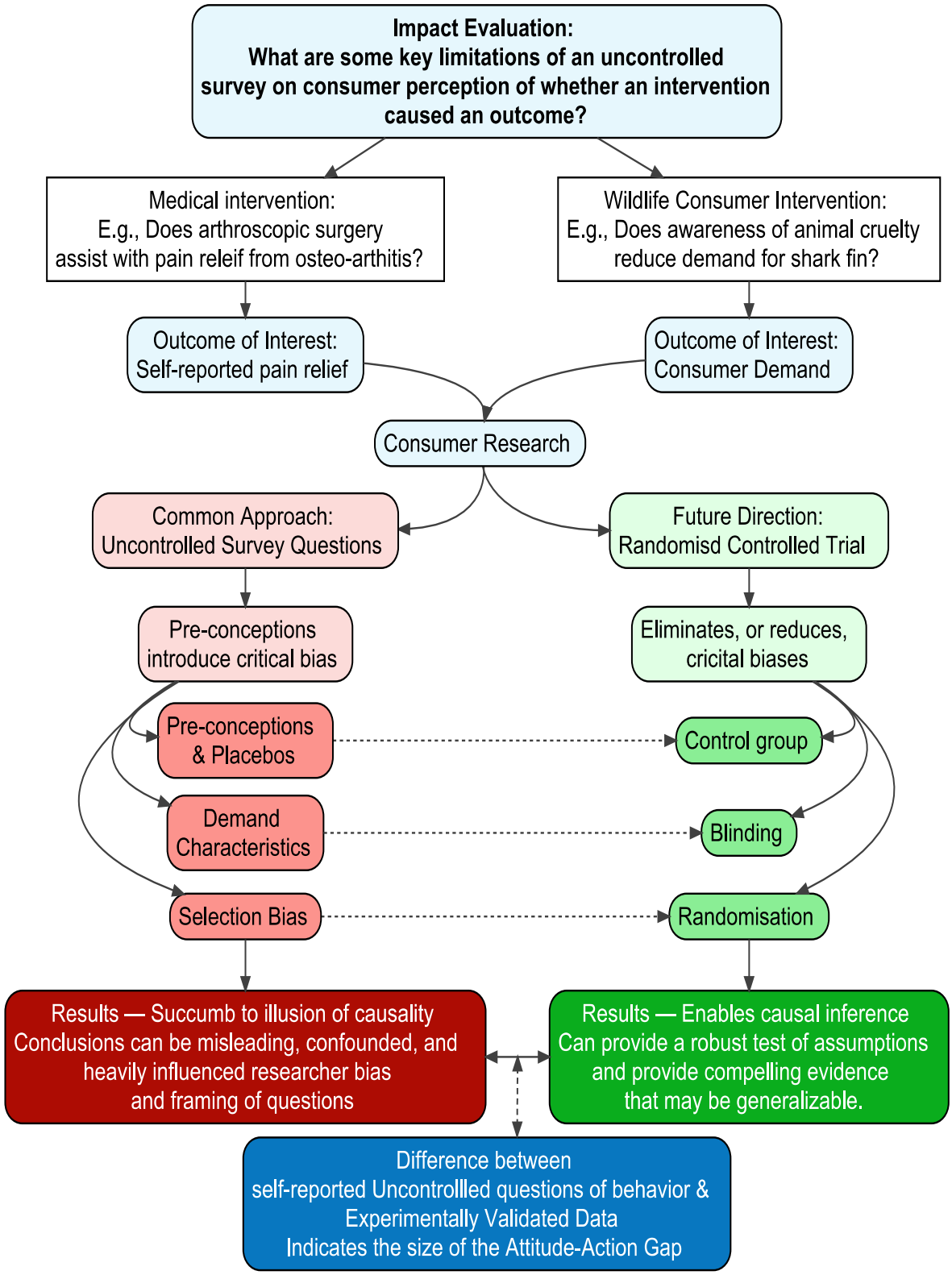

Figure 4. This figure provides a visual comparison between uncontrolled survey designs and experimental approaches

3 IMPLICATIONS FOR RESEARCH ON WILDLIFE CONSUMERS

3.1 Limitations of wildlife consumer perceptions surveys

To put the surgery example into the wildlife consumer context, there is often considerable difference between people's perception of which strategies are effective and empirical evidence of the same.

At present, uncontrolled consumer surveys (which ask consumers to select from a finite range of possible responses) are being used as part of formative research to uncover likely drivers of, and promising interventions to reducing, consumer demand.

However, uncontrolled surveys asking consumers which interventions will change their behaviour, can only indicate what participants believe will change their behaviour—they cannot provide evidence as to what interventions will actually change consumer behaviour.[3]

Further, uncontrolled survey questions can actually serve to foster misleading conclusions about intervention effectiveness through errors in causal inference (e.g., via causal illusions, demand characteristics, and selection biases).

Instead, researchers looking to conduct formative research into the underlying drivers of consumption would be better served by adopting qualitative research methods (e.g., open-ended questions, journey mapping, and focus groups; (Rare and The Behavioural Insights Team, 2019)).

Once researchers have an understanding of these drivers, they should then pilot test potential targeted intervention strategies (e.g., targeted messaging) using a robust experimental design and a representative sample of target consumers.

[3] Likewise, contrary to recent research claims (Cheung et al., 2018), asking experts (e.g., traditional remedy practitioners who prescribe remedies containing wildlife products) if a specific intervention (e.g., information about the lack of scientific evidence for a remedy) will change their behaviour (e.g., stop prescribing a wildlife-based remedy)—does not provide evidence on whether said intervention will actually change their behaviour, or not.

Box 5. "Path analysis" does not generate evidence of causality

In a recent uncontrolled survey of shark fin consumers, market researchers used path analysis (a.k.a., causal path modelling) to "uncover direct and indirect causal effects" that drive "the intention to consumer shark fin soup". However, many of the assertions made were misleading, because they relied on data obtained using an uncontrolled consumer survey, a method that does not facilitate causal inference. Path analysis is a hybrid between regression analysis and structural equation modelling. Such methods enable researchers to describe the relationships between numerous underlying factors. For example, they can be used to describe the strength of relationships between the perceived health benefit, perceived symbolic value, and the intention to consume shark fin soup.

Importantly, the research design used to obtain the data used in path analysis determines what statements can be made regarding the relationships depicted in causal path models. ***Uncontrolled survey designs cannot be used to make claims about causality.*** *Instead, uncontrolled surveys data can only be used to make speculative, and tentative accepted, assertions about causality, until such relationship can be experimentally tested. In contrast, experimental designs (and often to a lesser extent, quasi-experimental and longitudinal designs) do permit claims about causality—provided that potential confounding factors are properly controlled (Bozionelos, 2003).*

To determine the effectiveness of an intervention, small randomised controlled trials are usually the easiest and most robust form of experimental design (see Figure 4). Randomised controlled trials aim to compare the behaviour of two (or more) effectively similar groups of people, where the only significant difference between the groups is the intervention being tested. This method allows researchers to infer that any differences in observed behaviour were most likely to be caused by the intervention. RCTs are considered by many to be the gold standard for testing causality.

3.2 When an RCT is not feasible, it may be possible to simulate a "natural experiment"

To determine what action(s) caused a particular change in behaviour (either directly observed or self-reported), alternatives need to be "eliminated", however plausible they are as explanations for change. In certain cases where it may not be feasible to run a controlled experiment (e.g., it would be unethical, or impractical, to only allocate a treatment to half a target population) then we might be able to simulate a "natural experiment", whereby certain pieces of data can be used to eliminate all but one (or a few) explanations for an observed

change.[4] For example, data showing that the price of a product had not changed, would help exclude the price of a product as a plausible explanation for an observed change in consumer behaviour.

Natural experiments do however rely on having access to abundant and reliable data. Indeed, such comprehensive datasets enable public health researchers to have compellingly established that smoking kills (Centers for Disease & Prevention, 2008) and that the MMR vaccine does not cause autism (Flaherty, 2011). Unfortunately, due to the illicit nature of the illegal wildlife trade, data on the consumption of many wildlife products is often patchy or non-existent, thus diminishing our ability to make robust causal inferences.

4 OBJECTIVELY MEASURING CONSUMER DEMAND:

4.1 How can we reliable measure wildlife consumer demand?

As considered further in the forthcoming Journal Article under Activity 5.5.5, the first aspect to consider around ensuring effective social science survey design, is selecting an appropriate indicator (or indicators)[5].

In the previous medical example, the indicator was the amount of pain reported by patients. In the context of DR initiatives, the most important measurement is likely to be real-world consumer demand (intent to purchase and purchase) of a particular illegal wildlife product.

To measure the effectiveness of a DR initiative therefore, ideally social science survey indicators would focus on quantifiable observations or self-reported opinion based data, such as the amount of a particular wildlife product that a population consumes and/or the average price that they paid for that product.

[4] For assistance on establishing which methods are most appropriate for evaluating a particular conservation intervention, practitioners should refer to existing guidance, such as research decision trees, which have been specifically developed for evaluating conservation interventions (Woodhouse, de Lange, & Milner-Gulland, 2016).

[5] Research on consumer intent to purchase has several limitations not least because intention is poor indicator of our actual future behaviour. Indeed, many fluctuating factors (price, social pressure, opportunity, temptation etc) will influence whether or not people will actually stick to their stated intentions, so there is often a substantial gap between people's expressed intention and their actual behaviour (a.k.a., the intention-action gap).

In the real world, such data is challenging to obtain, without the use of sensitive questioning techniques or experimental methods, as introduced in the Discussion Paper for Activity 5.5.3&4. Innovative measurement tools (such as sensitive questioning techniques) may therefore be required.

4.2 Tools for measuring illicit behaviours

There are many innovative measurements tools that can help reveal people's non-reported uses of, or underlying attitudes towards, illicit products. These techniques, described in detail elsewhere (see the 'Options Paper' produced by Walsh & Vogt, 2019, under Activity 5.5.1) include: field observations, unmatched count techniques, implicit association tests, and computerised data collection.

Deciding which measurement tool is most appropriate for assessing demand will depend on the given product, its geographical and cultural context, and other practical considerations (e.g., such ethics, resources, ability to recruit target consumers). As each method has its own strengths and weaknesses, researchers should triangulate the datasets arising with other more 'observation' based, rather than 'opinion' based, data.

Ideally, multiple measurement tools would be tested simultaneously, so that the results can be statistically compared (i.e., cross-validated). A measurement tool is often said to be validated, when multiple tools return relatively consistent results (e.g., have similar effect strength and direction), which suggests that aspects of the same underlying consumer demand is being measured. Aspects such as this will be considered in more detail in the forthcoming Journal Article under Activity 5.5.5.

5 CONCLUSION:

5.1 Moving towards meaningful evaluations of demand-reduction interventions for wildlife consumers.

This Technical Briefing Paper has explored some of the considerations relevant to conducting high quality social science surveys, which form one of the three datasets proposed under the Discussion Paper for Activity 5.5.3 and 5.5.4.

Using an exploration of case studies from the field of medicine, where significant effort has been invested and insight gained, into how to strengthen impact evaluation processes.

This has helped to highlight the shortcomings of a non-experimental approach and associated with this, three psychological biases that could impair efforts to gather meaningful insights;

- The illusion of causality,
- Demand characteristics, and
- Selection bias.

These three key biases help to explain why, for example, a recent review of DR initiatives (which assessed some 236 demand reduction campaigns) concluded that the *"lack of robust impact evaluation made it difficult to draw insights to inform future efforts"*.

Importantly, this paper has not: addressed numerous other biases that can undermine research evaluations (for further guidance see (Moher et al., 2010); outlined a broader process for designing evidence-based interventions (see McKenzie-Mohr & Schultz, 2014; MacFarlane et al., 2020); nor addressed the many other aspects of research and collaboration that are key to effectively addressing illicit wildlife trades (e.g., recruiting target consumers for research or obtaining county-specific research permits).

Instead, this paper has attempted to complement existing work in the demand reduction space (Milner-Gulland et al., 2018; Rare and The Behavioural Insights Team, 2019; Burgess et al., 2018; TRAFFIC, 2018), and to provide a platform for the Journal Article that will follow it.

References

Barberia, I., Blanco, F., Cubillas, C. P., & Matute, H. (2013). Implementation and assessment of an intervention to debias adolescents against causal illusions. *PLoS One, 8*(8), e71303.

Becker, G. M., DeGroot, M. H., & Marschak, J. (1964). Measuring utility by a single-response sequential method. *Behav Sci, 9*(3), 226-232. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/5888778

Bozionelos, N. (2003). Causal path modeling: what it does and what it does not tell us. *Career development international, 8*(1), 5-11.

Breidert, C., Hahsler, M., & Reutterer, T. (2006). A review of methods for measuring willingess-to-pay. *Innovative Marketing, 2*(4), 8-32.

Burgess, G., Zain, S., Milner-Gulland, E. J., Eisingerich, A. B., Sharif, V., Ibbett, H., . . . Sohl, H. (2018). *Reducing demand for illegal wildlife products: Research analysis on strategies to change illegal wildlife product consumer behaviour*. Retrieved from

Centers for Disease, C., & Prevention. (2008). Smoking-attributable mortality, years of potential life lost, and productivity losses--United States, 2000-2004. *MMWR. Morbidity and mortality weekly report, 57*(45), 1226.

Cheung, H., Mazerolle, L., Possingham, H. P., & Biggs, D. (2018). Medicinal Use and Legalized Trade of Rhinoceros Horn From the Perspective of Traditional Chinese Medicine Practitioners in Hong Kong. *Tropical Conservation Science, 11*, 1940082918787428.

CITES (2017) 69th meeting of the Standing Committee SC69 Doc. 15. Demand Reduction. https://cites.org/sites/default/files/eng/com/sc/69/E-SC69-15.pdf

Crichton, F., & Petrie, K. J. (2015). Health complaints and wind turbines: The efficacy of explaining the nocebo response to reduce symptom reporting. *Environ Res, 140*, 449-455. doi:10.1016/j.envres.2015.04.016

de Barra, M. (2017). Reporting bias inflates the reputation of medical treatments: A comparison of outcomes in clinical trials and online product reviews. *Social Science & Medicine, 177*, 248-255. doi:https://doi.org/10.1016/j.socscimed.2017.01.033

Flaherty, D. K. (2011). The vaccine-autism connection: a public health crisis caused by unethical medical practices and fraudulent science. *Annals of Pharmacotherapy, 45*(10), 1302-1304. doi:10.1345/aph.1Q318

Hemila, H., & Chalker, E. (2013). Vitamin C for preventing and treating the common cold. *Cochrane Database Syst Rev*(1), CD000980. doi:10.1002/14651858.CD000980.pub4

Ioannidis, J. P. A. (2017). Does evidence-based hearsay determine the use of medical treatments? *Social Science & Medicine, 177*, 256-258. doi:10.1016/j.socscimed.2017.02.004

Jonas, W. B., Crawford, C., Colloca, L., Kaptchuk, T. J., Moseley, B., Miller, F. G., . . . Meissner, K. (2015). To what extent are surgery and invasive procedures effective beyond a placebo response? A systematic review with meta-analysis of randomised, sham controlled trials. *Bmj Open, 5*(12), e009655.

Lilienfeld, S. O., Ritschel, L. A., Lynn, S. J., Cautin, R. L., & Latzman, R. D. (2014). Why Ineffective Psychotherapies Appear to Work: A Taxonomy of Causes of Spurious Therapeutic Effectiveness. *Perspectives on Psychological Science, 9*(4), 355-387. doi:10.1177/1745691614535216

MacFarlane, D., Hurlstone, M. J., & Ecker, U. K. H. (2018a). Reducing Demand for Ineffective Health Remedies: Overcoming the Illusion of Causality. *Psychology & Health, 33*(12), 1472-1489. doi:10.1080/08870446.2018.1508685

MacFarlane, D., Hurlstone, M. J., & Ecker, U. K. H. (2018b). Reducing Demand for Ineffective Health Remedies: Overcoming the Illusion of Causality [Supplemental Material]. *Psychology & Health*. doi:10.1080/08870446.2018.1508685

MacFarlane, D., Hurlstone, M. J., & Ecker, U. K. H. (2020). Protecting consumers from fraudulent health claims: A taxonomy of psychological drivers, interventions, barriers, and treatments. *In press for Social Science and Medicine*

Matute, H., Yarritu, I., & Vadillo, M. A. (2011). Illusions of causality at the heart of pseudoscience. *Br J Psychol, 102*(3), 392-405. doi:10.1348/000712610X532210

McKenzie-Mohr, D., & Schultz, P. W. (2014). Choosing effective behavior change tools. *Social Marketing Quarterly, 20*(1), 35-46. doi:10.1177/1524500413519257

Milner-Gulland, E. J., Cugnière, L., Hinsley, A., Phelps, J., Sas-Rolfes, t., & Verissimo, D. (2018). *Evidence to Action: Research to address the illegal wildlife trade. Briefing note to policy-makers and practitioners.* Retrieved from

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gotzsche, P. C., Devereaux, P. J., . . . Altman, D. G. (2010). CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ, 340*, c869. doi:10.1136/bmj.c869

Moseley, J. B., O'Malley, K., Petersen, N. J., Menke, T. J., Brody, B. A., Kuykendall, D. H., . . . Wray, N. P. (2002). A Controlled Trial of Arthroscopic Surgery for Osteoarthritis of the Knee. *New England Journal of Medicine, 347*(2), 81-88. doi:10.1056/NEJMoa013259

Rare and The Behavioural Insights Team. (2019). *Behavior Change For Nature: A Behavioral Science Toolkit for Practitioners.* Retrieved from Arlington, VA:

Reinhart, A. (2015). Statistics done wrong: The woefully complete guide. In (pp. 39-54): No starch press.

Siemieniuk, R. A. C., Harris, I. A., Agoritsas, T., Poolman, R. W., Brignardello-Petersen, R., Van de Velde, S., . . . Kristiansen, A. (2017). Arthroscopic surgery for degenerative knee arthritis and meniscal tears: a clinical practice guideline. *BMJ, 357*, j1982. doi:10.1136/bmj.j1982

Stewart-Williams, S. (2004). The placebo puzzle: putting together the pieces. *Health Psychol, 23*(2), 198-206. doi:10.1037/0278-6133.23.2.198

Sunstein, C. R. (2019). *How Change Happens*: Mit Press.

Sutherland, W. J., Taylor, N. G., MacFarlane, D., Amano, T., Christie, A. P., Dicks, L. V., . . . Ockendon, N. (2019). Building a tool to overcome barriers in research-implementation spaces: The conservation evidence database. *Biological Conservation, 238*, 108199.

Thorlund, J. B., Juhl, C. B., Roos, E. M., & Lohmander, L. S. (2015). Arthroscopic surgery for degenerative knee: systematic review and meta-analysis of benefits and harms. *British Journal of Sports Medicine, 49*(19), 1229. doi:10.1136/bjsports-2015-h2747rep

Thrasher, J. F., Rousu, M. C., Hammond, D., Navarro, A., & Corrigan, J. R. (2011). Estimating the impact of pictorial health warnings and "plain" cigarette packaging: evidence from experimental auctions among adult smokers in the United States. *Health Policy, 102*(1), 41-48. doi:10.1016/j.healthpol.2011.06.003

TRAFFIC. (2018). *Monitoring and evaluating behaviour change amongst illegal wildlife product consumers: Good practice guidlines*. Retrieved from https://www.traffic.org/site/assets/files/1851/sbcc-me-good-practice-guidelines-2018.pdf

USAID. (2017). *Measuring efforts to combat wildlife crime: A toolkit for improving action and accountability.* Retrieved from https://rmportal.net/biodiversityconservation-gateway/resources/archived-projects/measuring-impact/mi-project-

resources/measuring-efforts-to-combat-wildlife-crime-a-toolkit-for-improving-action-and-accountability-pdf

(2019, April 29). *All the World's a Stage—Including the Doctor's Office* [Retrieved from https://www.npr.org/2019/04/29/718227789/all-the-worlds-a-stage-including-the-doctor-s-office

Veríssimo, D., & Wan, A. K. Y. (2019). Characterizing efforts to reduce consumer demand for wildlife products. *Conservation Biology, 33*(3), 623-633.

Walsh, J., & Vogt, S. (2019). *Strengthening Demand Reduction Measurement: Options on methods from behavioural science (unpublished)*. Retrieved from

Woodhouse, E., de Lange, E., & Milner-Gulland, E. J. (2016). *Evaluating the impacts of conservation interventions on human wellbeing. Guidance for practitioners.* Retrieved from London, Easthope.

Disclaimers

The EU is not responsible for the contents of communication materials prepared by implementing partners, all of which must include a standard disclaimer, translated into (the) local language(s) where appropriate.