Against methodological gambling[*]

This is the accepted version of a paper that is forthcoming in Erkenntnis.

Please cite the published version.

Borut Trpin

Ludwig-Maximilians-Universität München
Munich Center for Mathematical Philosophy
borut.trpin@lrz.uni-muenchen.de

Abstract

Should a scientist rely on methodological triangulation? Heesen et al. (2019) recently provided a convincing affirmative answer. However, their approach requires belief gambles if the evidence is discordant. We instead propose epistemically modest triangulation (EMT), according to which one should withhold judgement in such cases. We show that for a scientist in a methodologically diffident situation the expected utility of EMT is greater than that of Heesen et al.'s (2019) triangulation or that of using a single method. We also show that EMT is more appropriate for increasing epistemic trust in science. In short: triangulate, but do not gamble with evidence.

1 Introduction

Social scientists often combine multiple methods in their research. Take, for instance, the hypothesis that parental resources do not affect the economic position of adult children with a college degree ("a college degree as the great equalizer" hypothesis; e.g., Hout, 1984, 1988; Torche, 2011). To test this hypothesis in practice a scientist needs to have a reliable way of measuring individuals' economic position. Torche (2011) points out that we may measure economic status

[*] Special thanks to Mariangela Zoe Cocchiaro for helping me clarify ideas in this paper. Thanks also to two anonymous reviewers for their valuable comments on earlier versions of the paper. The research was supported by Alexander von Humboldt Foundation.

with at least four indicators: social class, occupational status, individual earning, and total family income. The indicators each have their pros and cons. What is, then, the most reasonable way to determine one's economic status? Plausibly, we should gather as much evidence as possible from multiple methods. The problem is more difficult, however, if evidence is discordant, as it will often be if we are to use multiple indicators which "will provide a dissimilar evaluation" (Torche, 2011, p. 774). As an example, suppose that three indicators suggest high economic status, but the individual's earning is very low. It seems that we can still safely conclude that the individual is rich and find a plausible explanation of low earning.

The principle that one should not pursue only one line of evidence (e.g., individual earning when estimating economic status) is common in scientific research. We should pursue multiple methods and thus reduce the risk of using an unreliable method. Similarly, we do not want to replicate an unreliable experiment because even if it successfully replicates, the results would be misleading (see, e.g., Munafò and Smith, 2018).

To put it more precisely, the idea is that if evidence in favour of some hypothesis h is more varied, it provides better support for h. This position is known as the Variety of Evidence Thesis (VET) and has a number of defenders both among philosophers of science and practising scientists. Surprisingly, when we model these situations in a Bayesian framework, VET turns out to be false or at least limited in scope (Bovens and Hartmann, 2003; Claveau, 2013; Claveau and Grenier, 2019; for a non-Bayesian critique of VET see, e.g., Hudson, 2014).

In contrast, Heesen et al. (2019) recently proposed an approach according to which VET holds, the so-called Du Boisian methodological triangulation (DMT). They argue that if we have no reason to prefer one method to another, triangulation on the basis of results provided by multiple methods serves as a better guide to truth than the results of a single method. We find their conclusion plausible if evidence is concordant (see Stegenga, 2012, for a similar position). If it is not, however, their method simply instructs us to randomly endorse one of the results. This is intuitively problematic and, as we will show, arguably not a smart move.

In what follows, we will take a more detailed look at DMT (sections 2 and 3), where we will explain its advantages and disadvantages. We will then turn to our proposed epistemically modest triangulation (EMT, section 4), which overcomes some of the problems of DMT. We will then demonstrate that EMT has a higher expected utility than DMT (section 5) and discuss its advantages as well as some potential objections (section 6). Finally, we will take stock of EMT in Conclusion (section 7).

2 Du Boisian methodological triangulation

A common problem in science, and perhaps even more so in social sciences specifically, is that there are multiple methods available to investigate most phenomena. At the same time it is often not clear which method is the most appropriate for a given research question. W.E.B. Du Bois, an American sociologist best known for his research of race-related issues, experienced this problem when he tried to find the answer to the question "How much do Negros earn?" (Du Bois, 1899; cited from Heesen et al., 2019). He estimated the answer on the basis of multiple methods (visual appearance of the homes, direct responses to the question, etc.) and argued that this pluralistic approach helped him overcome the limitations of each particular method. A similar multi-method approach is still commonplace in social sciences (e.g., the use of multiple indicators to estimate economic status; see DiPrete, 2020, p. 384-386, for a review). Although Du Bois remains relatively ambiguous about his triangulation on the basis of multiple methods, Heesen et al. (2019) provide a mathematical formalisation of his approach. They call it Du Boisian methodological triangulation (DMT).

DMT is formulated in analogy to elections. We can imagine that each method is in a sense a voter. The set of potential answers provided by the methods is then analogous to the candidates that run in an election. Each method provides one answer (has one "vote"). DMT then suggests that we should embrace the most common answer (viz. the candidate with the plurality of votes). For instance, suppose we have three methods to estimate an individual's economic standing: occupational status, individual earning and total family income. Suppose the methods can only provide three mutually exclusive answers – A: low status, B: middle status, C: high status. If all methods return ("vote for") the same answer, say A, then it is uncontroversial that we have reasons to embrace answer A, i.e., that the individual's economic status is low. Triangulation does not add much in such a case because we would reach the same answer by using any of the three methods alone, except that A seems to be better supported because it was obtained by multiple methods.

Now suppose that two methods provide answer A and one B. DMT then suggests that we should still embrace A. This is intuitively plausible because we assume the methods are more likely to provide a true than a false answer – otherwise it would be unwise to use them. More importantly, all we need to assume is that at least one of the methods at least weakly correlates with the truth (performs better than a random guess). DMT will then be more likely to provide a true answer than randomly selecting one of the methods and endorsing its results. The motivation for this is that scientists often do not know which method is the most likely to deliver a true result. This does not mean that they should just select one of the methods and stick with it. Instead, they should triangulate if they find themselves in such states of methodological diffidence because DMT will be more likely to provide a true result (see Heesen et al., 2019, for a proof).

While this is good news for triangulators, DMT comes with two caveats. As we mentioned, DMT is appropriate when an agent is in the position of methodological diffidence, that is, when no method is known to be the most accurate. Such situations might be commonplace, but it also seems that there will often be reasons to prefer some methods. For instance, social status is an aggregated indicator which combines a number of aspects, while individual earning is not, so it is possible that the former would be more reliable for estimating economic status, *ceteris paribus*. An objection could therefore be raised that the methods should be weighted in relation to their past performance. To continue with the electoral analogy, the situation should be akin to a weighted vote where more weight is assigned to more reliable methods. However, if one of the methods tracks truth better, then it is no longer clear whether triangulation is still necessary. A definitive resolution of this objection would exceed the scope of the present paper, so we (like Heesen et al., 2019) limit ourselves to situations where an agent has no reason to prefer one method to another.[1] After all, such situations arguably occur and if we can show that triangulation is useful at least in such cases, this provides an argument in its favour.

The second problem of DMT is more worrying. So far we have talked about the cases where the plurality of (or all) methods provide the same answer. For instance, the answers A, A, B by methods 1, 2, and 3, respectively. However, if no unique answer wins the plurality of "votes", DMT suggests that we should simply take a gamble.

By means of an example: suppose that we are trying to estimate an individual's economic standing and we only take note of her occupational status (low) and her total family income (high). Suppose also that there is no reason to prefer either method: there are individuals who are economically well-off due to high family income despite low occupational status, as well as individuals who are not well-off due to their low occupational status and despite high total family income. It is not possible to reliably triangulate on the basis of such evidence, right? At least according to DMT it is. We should simply flip a coin and conclude whether the individual is well-off or not.

Such an approach might make sense in political elections. If there is a perfect tie among two or more candidates, the final call can be made by a random choice. But with a large number of votes in a typical election, such situations happen rarely and it is not obvious that only one candidate is correct (in whichever sense).[2] With methodological triangulation this does not

[1] For some issues related to weighted judgement aggregation see, e.g., Klein and Sprenger, 2015; Boyer-Kassem, 2019, and Martini and Sprenger, 2017 for a survey. Thanks to an anonymous reviewer for this suggestion.

[2] See, however, Wiblin (2020) for a recent discussion of correct/wrong votes and the role of epistemic modesty in democratic elections.

hold: "ties" (no unique plurality result) can happen quite often because the number of methods in practice is often low (e.g., below 5 or 10 methods). Furthermore, if the answers are jointly exhaustive and mutually exclusive, then only one answer is true and the others are false. We find such gambles troubling and will suggest that in cases like this a triangulator should rather withhold judgement and either seek further evidence or reasons to favour some of the methods. Putting our worries and the discussion of what withholding judgement would mean in practice aside for a moment, does randomisation even play any vital role in the justification of DMT? As we will now show, its advantages simply cannot be proven without it.

3 Methodological gambling

We will start with a simple toy example. Suppose a social scientist is analysing the relationship between poverty and health. Let us focus only on the latter – how may she determine the health status of an individual? Let us further suppose three independent methods M_1, M_2, and M_3 may each provide three mutually exclusive and jointly exhaustive answers A, B or C to this question. The methods could be, for instance, self-report health assessments (M_1), an aggregate measure based on the incidence of specific chronic illnesses (M_2), and an aggregate measure based on anthropometric indicators like body mass index and blood pressure (M_3).[3] Suppose the answers provided by A, B, and C are consistent with good, average, and poor health, respectively. We further assume that there is no reason to prefer either of the methods.

Without any loss of generality, assume that the correct answer is A. The worst case scenario for triangulation would be if one of the methods is perfect in the sense that it always gives the true result, while the other methods are no better than guesswork. In such a case, triangulation will perform worse than the best method. Suppose that for method M_1, $P_{M_1}(A) = 1$; the probability that M_1 gives answer A is 1 (100%). The other two methods, M_2 and M_3, are just glorified guesswork, so their probability to produce any result, including the true one, is $P_{M_2}(A) = P_{M_3}(A) = 1/3$ (1/number of possible answers). If we know which method is perfect, then obviously M_1 will perform best as it always gives the true result. But if we do not know which method is the best and randomly select one of them – we call this approach diffident purism –, then the probability of getting the correct answer equals $1/3 \times 1 + 2/3 \times 1/3 = 15/27$. DMT, on the other hand, will produce the correct answer with probability $17/27$, as we show below (for a generalisation see Heesen et al., 2019, Theorems 2 and 3).

To understand why DMT has a higher probability to return the true result A, we need to understand how DMT functions. There are three ways it will provide the correct result:

[3] See Aaberge and Brandolini (2015) for a review of additional methods related to well-being that are used in poverty studies.

(i) When all methods provide the correct result,

(ii) When the plurality of methods provide the correct result,

(iii) By randomisation when there is no unique plurality answer and one of the tied answers is correct.

In cases (i) and (ii) the situation is simple: the correct answer is given by the plurality of methods, so DMT endorses it. The remaining case (iii) is a bit trickier. Suppose multiple answers are in a tie and the correct answer is among them. In case of three methods with three possible answers like above this means that the methods each provide answers A, B, and C. If we randomly select one of the answers, the probability that we go with the correct one is 1/(number of tied answers). Here, 1/3.

To calculate the probability that DMT gives the true answer we then need to sum the probability of cases (i) and cases (ii). We then also add the probability of cases (iii) divided by the number of tied answers. In our example: $1/9 + 4 \times 1/9 + 2 \times (1/3) \times (1/9) = 17/27$.[4]

Importantly, if we would simply ignore cases (iii), then DMT*, as we denote this variant of DMT, would not necessarily outperform diffident purism. In the above example, both DMT* and diffident purism have the same probability of delivering the true result (15/27=5/9), although it is also possible that one or the other would be more likely to produce a true result in other situations.[5] In other words: if we leave gambling to casinos and remove it from methodological triangulation, DMT is no longer vindicated.

Note that a similar objection against triangulation on different grounds has already been raised by Stegenga (2012, p. 216-7). As he points out, triangulation might be valuable if evidence is not discordant. However, he provides evidence against the claim that "multimodal evidence is often concordant – /.../ this is an empirical claim which is false."

Nevertheless, one might still argue that we simply need to bite the bullet and accept that methodological gambling is the price one has to pay to be more likely to uncover true results – if, as noted, it is not clear which method is the most reliable. The idea could be that even if DMT is not ideal, it is the best we have. After all, as Skipper (2020) recently demonstrated, belief gambles in general seem irrational, but it is very hard to explain why exactly when we assess them more precisely (e.g., with epistemic decision theory). In what follows we will take on this challenge and show that we can do without methodological gambling and still outperform both diffident purism and DMT.

[4] M_1 always provides the true answer A. Hence, the probability of case (i): A,A,A by M_1, M_2, and M_3, respectively, is $1 \times 1/3 \times 1/3$. There are four possible outcomes for cases (ii): A,A,B; A,A,C; A,B,A; A,C,A, each with probability 1/9. Finally, there are two possible outcomes for cases (iii): A,B,C; A,C,B, each with probability 1/9. The correct answer might be randomly endorsed, so we multiply the probability of these cases by 1/3.

[5] For example, assume that one of the methods always provides the true answer and the others are just guesswork. Diffident purism outperforms DMT* if there are 2 methods with 2 possible answers, and DMT* outperforms purism if there are 3 methods with 2 possible answers. The proof is straight-forward, so we leave it as an exercise for the reader.

4 Epistemically modest triangulation

We suggest that a triangulator should not endorse any of the results when the evidence is inconclusive or discordant in the sense that there is no single prevalent result (obtained by the plurality of methods).[6] We call this approach epistemically modest triangulation (EMT) because it makes no pretence of having a privileged access to the truth when the evidence shows no clear inclination towards any of the potential results.[7]

We justify the move from DMT to EMT conceptually on the basis of the insights from the epistemology of disagreements. In case two epistemic peers disagree about p, it is rational for them to reduce their confidence in it (e.g., Christensen, 2007).[8] Because we are addressing situations of methodological diffidence in which it is not clear why any method should be considered to be more reliable than another, we can understand discordant methods as analogous to epistemic peers who find themselves in a disagreement. Hence, if we have no good reason to consider a specific method to be better than another and there is no answer obtained by the plurality of methods, then we need to reduce our confidence in the results.

What exactly is EMT then? It is a triangulation strategy that behaves in exactly the same way as DMT when the methods provide concordant results, i.e. when one of the results is obtained by the plurality of methods. A triangulator then endorses this prevalent result. However, if no specific result prevails (e.g., three methods each give a different answer), then EMT avoids the randomisation step and instead prescribes that the triangulator should remain epistemically modest and withhold her belief in any of the results. Before moving on we need to address two potential worries:

1. What does withholding belief mean for triangulation?

Let us return to our previous toy example of a scientist interested in health of some group (e.g., to investigate correlations between poverty and health). Suppose self-report assessments suggest the individuals in the group are in good health, while anthropometric indicators like blood pressure suggest poor health, and the incidence of specific chronic illnesses suggests average health. In this case, EMT prescribes the triangulating scientist not to conclude anything about the health of this group. In practice this means that in such scenarios evidence is discarded because it is inconclusive. This could be understood as a call for further research, an analysis of how reliable the methods are, use of additional methods, or something else that would help decide the situation. As long as the evidence is inconclusive, however, it is not possible to use

[6] Note that when we can further distinguish discordance in a graded sense. If we say the results are *fully* discordant, we mean that n methods provide n different answers.

[7] A reader might notice that EMT is just like DMT* except that instead of ignoring outcomes where no answer is provided by the plurality of methods it actively prescribes the triangulator to withhold judgement.

[8] We are aware that there are many versions of conciliatory as well as non-conciliatory positions and pluralist positions in the epistemology of disagreements. We address this question with respect to scientific disagreements elsewhere.

EMT to infer what the health status of the group under investigation is – hence, it is also not (yet) possible to include the evidence about this group in the analysis of the relationship between health and economic status. We believe this should not be particularly problematic as it also corresponds to one of the ways scientists deal with uncertainty in practice – problematic, inconclusive evidence is often simply avoided instead of resolved (see, e.g., Schickore and Hangel, 2019, p. 6). We do not argue that this is the best response when no result is obtained by the plurality of methods or that a resolution of these cases to obtain a prevalent answer is not to be preferred. We only want to make clear that inconclusive evidence is already often disregarded in practice. This can also lead to the so-called file-drawer problem (Rosenthal, 1979) – the situation in which negative or inconclusive results are under-reported and hence lead to biased syntheses of evidence. This is a well-known problem, so a number of techniques to detect and avoid it in, e.g., meta-analysis have already been developed (see Borenstein et al., 2011, Ch. 31, for an overview). Whether this saves potential issues of discarding inconclusive evidence or not remains an open question, but the fact that inconclusive outcomes are not ideal will be reflected in our formal justification of EMT in the following section, and we will still be able to show it outperforms DMT.

2. Could compounding results solve the issue in a more elegant way?

Note also that when we talk of discordance, we talk of discordance among specific (non-compound, "atomic") results. For instance, if 4 methods with 5 potential answers A, B, C, D return answers A, A, B, and B, respectively, EMT prescribes that we should withhold our judgement about the correct result. An alternative approach to triangulation might be feasible, which would preserve the spirit of EMT (avoiding errors if evidence is discordant) but which would at the same time also not discard any evidence. Let us call one such alternative approach compound methodological triangulation (CMT) to see why it is, in our opinion, not a feasible alternative to EMT.

CMT, which compounds answers, would be exactly like EMT and DMT when some specific answer is prevalent. If evidence is discordant, however, CMT instructs us to endorse a disjunction of all tied answers.[9] Specifically, if 4 methods provide answers A, A, B, B (out of A, B, C, D), CMT would thus lead to an endorsement of the disjunction $A \vee B$. If the answers were, instead, A, B, C, and D – what we earlier described as a fully discordant outcome (see Fn. 6) – CMT would endorse $A \vee B \vee C \vee D$, which is an answer that would not be false, but also completely uninformative.

At this stage we observe two points: first, it is hard to judge in which cases CMT could be useful in practice. When could a disjunction of answers help us? For example, suppose we are

[9] Thanks to an anonymous reviewer for this suggestion.

estimating economic status. It seems one relatively useful outcome is if we are able to determine that someone is either in a very low *or* low economic position. We can merge the two answers together (e.g., here that the individual is in a broadly low position). It is worth noting that this will not always, or perhaps even typically, work. On the other hand, a disjunction of low and high economic status does not seem to be a very insightful result. Second, it is not clear how to evaluate such disjunctions, both with respect to their probability as well as informativeness. It seems intuitively obvious that a disjunction of two results is more impressive if the number of potential results is high (e.g., 10) than when it is low (e.g., 2 of 3 possible answers in a disjunction). Moreover, if the disjunction contains all possible answers and the answers are jointly exhaustive, then the disjunction is a tautology and as such completely uninformative, although necessarily true. How could one then compare the performance of CMT to other triangulation techniques? Plausibly, CMT will have a higher degree of uncovering the truth than both EMT and DMT, but we then also need to have a measure of informativeness, which needs to be taken into account when comparing different approaches. Otherwise we might simply think of yet another triangulation approach: always infer to a disjunction of all results. Such an approach would not even be outperformed by perfect methods that always provide the true answer.

Putting these conceptual worries aside, even if we had a way of taking informativeness into account, we would still need additional machinery to evaluate different disjunctive answers because some disjunctions make more sense than others. This notion may perhaps be related to the coherence of evidence. The disjunctions whose disjuncts fit together well should presumably be in some way favoured to just any disjuncts that are tied. As we saw in the above example, it seems clear that we would prefer a disjunction of very low and low economic status to another of a very low and high status. We leave this question open for further research, not the least because the problems arise already when we need to decide how to measure coherence of propositions (the literature is rife with issues; see, e.g., Koscholke, 2019 for just one of the more recent examples).

Another issue that we need to also point out with respect to all considered triangulation approaches (DMT, EMT, CMT) is that we did not consider the nature of the answers provided in the triangulation. Instead, we just assumed the simplest categorical and mutually exclusive form of answers (e.g., "this individual's economic status is high") to establish the advantages of EMT over other approaches. A reasonable worry that one may raise is that different triangulation strategies might also differ in their performance when answers are numerical and commensurable, probabilistic, multi-valued, and so on.[10]

It is not clear how EMT would perform in such cases. Suppose we have two methods which

[10] Thanks to an anonymous reviewer for pointing out this option.

provide probabilistic answers to a question like "What is the probability that in some specific group an individual's income falls into some specific range?" It is quite unlikely that the two answers will be exactly the same, so EMT will typically lead to withheld judgement, especially if we use multiple methods. CMT, on the other hand, leads to a set of specific values which may also be useful for further analysis. It may seem, then, that CMT might be more appropriate at least in such cases.

This is not necessarily bad news for EMT for two reasons. First, we may limit the application of EMT only to cases where categorical and mutually exclusive answers may be provided – and such cases are common in scientific practice. If it outperforms DMT, CMT, and diffident purism in this context, this already suffices to show at least its relative advantage. Second, similarly as CMT needs to be developed further to overcome the issues sketched above, related to informativeness and coherence of the disjuncts, EMT may as well in some sensible way be adapted so that it would successfully aggregate non-categorical answers (e.g., by providing overarching categorical answers that subsume more precise answers). We leave such investigations for further work.

CMT, then, turns out to be an intuitively interesting alternative, but it opens a large number of questions that exceed the scope of this paper. EMT, on the other hand, is manageable and with the inclusion of withheld beliefs also corresponds to scientific practice at least descriptively, as we have shown above. The more important question is whether there are also normative reasons to prefer withholding judgement in some cases.

We have already seen that DMT outperforms diffident purism exactly because of the step we want to avoid, i.e. methodological gambling. Indeed, it immediately follows that EMT will be less likely to produce a true result if it delivers no result when the methods are discordant. No risk, no gain, as the saying goes. But there are two sides to every coin – methodological gambling increases the probability that triangulation delivers a true result, while it at the same time also increases the probability of endorsing a false result.

In fact, methodological gambling implies that the probability of randomly selecting a false result is at least one in two. This happens in case two answers are tied and one of them is true and the other false. As soon as more than just two results are tied (e.g., three methods with three different answers), it is immediately more likely that triangulation of this kind will endorse a false result. Similarly, if the true result is not even among the tied ones, methodological gambling will certainly lead to a false result. Hence, the gambling component of DMT increases the probability of striking upon a true result, but it also makes it more likely that a false result will be triangulated upon when evidence is discordant. Our proposed epistemically modest triangulation takes this into account. An EMT-triangulator will be, therefore, less likely to randomly uncover truth, but she will also be less likely to believe in a falsehood. As scientists ought not only care about uncovering truth, but ought also avoid spreading false claims, we believe that this provides a good reason to use EMT instead of DMT. Besides these general reasons in favour of EMT, we will now turn to a proof that EMT outperforms both DMT and diffident purism under a reasonable assumption of the values assigned to believing a true or false result or withholding a belief.

5 Expected utility of EMT

Why should triangulators then not gamble with evidence if this is a sure way to improve the probability of uncovering true hypotheses both compared to diffident purism and EMT? The answer is simple, as we have already mentioned: because we also increase the probability of endorsing false results. Here we are assuming that endorsing truth is more valuable than endorsing falsehood. When we add the possibility of withholding judgement – as in the case of EMT – we assume that this option is not great, not terrible. Its value lies between the values assigned to believing in falsehood and truth. Formally,

$$u(F) < u(W) < u(T) \tag{1}$$

where $u(T)$, $u(W)$, and $u(F)$ represent the utility of endorsing a true claim, withholding judgement, and endorsing a false claim, respectively.[11] So much should be relatively uncontroversial (for the same point see also Wilholt, 2013, p. 237).

A potential objection could nevertheless be raised that the value of withholding judgement may in some cases be equivalent to that of endorsing a false claim, particularly if the results obtained by EMT are used in practical decision-making where there are just two options: to act or not to act. Withdrawing belief could in practice be equivalent to not acting, hence its value would not be strictly greater than that of believing a false claim.

This challenge can be addressed by simply making clear that we are interested in epistemic performance of EMT (see also Frances and Matheson, 2019, 2, for a more detailed discussion of differences between belief- and action-disagreements). In addition, it is important to keep in mind that decision-making takes place at a later stage – once EMT had already been used and, hence, should be evaluated separately from EMT. This is because if no result is obtained by EMT, the decision-maker simply acts independently of it, which does not mean that no action will be taken – EMT will merely not provide additional reasons to act or not to act.

[11] We interchangeably refer to believing and endorsing a claim or a result because we assume that the triangulator will only (publicly) endorse some claim if she believes it. Other more complicated cases of the interplay between beliefs and endorsements are outside of the scope of this paper (for more details see Fleisher, 2018, 2019).

For instance, suppose a scientist uses EMT to estimate population health in some region. The evidence is inconclusive, so she withholds belief and does not come to any conclusion. A policy-maker who makes her decisions with respect to the health status of this population then acts as if this inconclusive EMT study was never conducted. This is, of course, not ideal, but it demonstrates why it is important to distinguish between decision- and belief-contexts. We will return to some further aspects of this matter in the next section.[12]

Let us now start our proof that EMT is (epistemically) more valuable than both DMT and diffident purism with a special case. Suppose we assume that withholding judgement has neither a positive nor a negative value and that believing a true claim is just as valuable as believing a false claim is invaluable:

$$u(W) = 0 \tag{2}$$

$$u(F) = -u(T) \tag{3}$$

We already know that DMT is at least as likely to uncover truth as diffident purism (Heesen et al., 2019). Hence, DMT also has a higher expected utility:

$$\Pr_{DMT}(T)u(T) + \Pr_{DMT}(F)u(F) \geq \Pr_{DIFF}(T)u(T) + \Pr_{DIFF}(F)u(F) \tag{4}$$

where $\Pr_{DMT}(T)$ and $\Pr_{DIFF}(T)$ represent the probability of endorsing a true result for DMT and diffident purism, respectively, and likewise for false results (represented by F). This follows straight-forwardly because of the inequality (1) and the fact proven by Heesen et al. (2019): $\Pr_{DMT}(T) \geq \Pr_{DIFF}(T)$. Note that $\Pr(F) = 1 - \Pr(T)$ for both approaches.

To show that EMT dominates both DMT and diffident purism it therefore suffices to prove that it dominates DMT. We have already mentioned that EMT operates in the same manner as DMT when one of the answers is obtained by the plurality of methods. By means of an illustration, suppose there are two methods that can only produce two answers, a simple "yes" or "no", whether women are paid less than men in some industry. Suppose the correct answer is "yes". We then have the following four potential outcomes:

Outcome	Method 1	Method 2
1	yes	yes
2	yes	no
3	no	yes
4	no	no

[12] Thanks to an anonymous reviewer for pointing out this issue.

Both methods will deliver the same result in outcomes 1 and 4; the correct answer in the former and the false in the latter case. They diverge in outcomes 2 and 3. DMT advises to flip a coin, so there is a probability of 1/2 that we end up endorsing the true answer. EMT, in contrast, suggests we should withhold our judgement. We can then analyse the expected utility of the two approaches by splitting the possible outcomes, in general, into those where one of the answers dominates (i.e., it is provided by the plurality of methods), and those where none does (non-dominant cases; here cases 2 and 3). The expected utility of DMT is then:

$$\begin{gathered} EU(DMT) = EU(DMT, dom) + EU(DMT, \overline{dom}) = \\ \Pr_{DMT}(T, dom)u(T) + \Pr_{DMT}(F, dom)u(F) + \Pr_{DMT}(T, \overline{dom})u(T) + \Pr_{DMT}(F, \overline{dom})u(F) \end{gathered} \tag{5}$$

where dom and $\overline{dom}$ represent dominant and non-dominant outcomes. The expected utility of EMT, on the other hand, also includes the cases where judgement is withheld:

$$EU(EMT) = \Pr_{EMT}(T)u(T) + \Pr_{EMT}(F)u(F) + \Pr_{EMT}(W)u(W) \tag{6}$$

Note that DMT and EMT provide the same answers in dominant cases (above, cases 1 and 4), hence $\Pr_{EMT}(T) = \Pr_{DMT}(T, dom)$ and $\Pr_{EMT}(F) = \Pr_{DMT}(F, dom)$. It, therefore, suffices to show that EMT has a higher expected value in non-dominant cases to conclude that it has a higher expected value in general. That is, we need to show the following:

$$EU(EMT, \overline{dom}) \geq EU(DMT, \overline{dom}) \tag{7}$$

We make a reductio assumption that the inequality (7) does not hold. Keeping in mind that the probability that EMT suggests an agent should withhold judgement is exactly the same as the probability that DMT provides either a true or false result in non-dominant cases, we get:

$$\begin{aligned} EU(EMT, \overline{dom}) &< EU(DMT, \overline{dom}) \\ \left(\Pr_{DMT}(T, \overline{dom}) + \Pr_{DMT}(F, \overline{dom})\right) u(W) &< \Pr_{DMT}(T, \overline{dom})u(T) + \Pr_{DMT}(F, \overline{dom})u(F) \end{aligned} \tag{8}$$

Taking inequality (1) and equation (3) into account, we find the following after some rearrangement (the result is a simple corollary of A.2; see Appendix):

$$u(W) < 0 \tag{9}$$

But inequality (9) contradicts equation (2) we assumed above ($u(W) = 0$). Our assumption leads to a contradiction, so we conclude with:

$$EU(EMT, \overline{dom}) \geq EU(DMT, \overline{dom}) \tag{10}$$

and consequently $EU(EMT) \geq EU(DMT) \geq EU(DIFF)$. Moreover, if we assume that at most one method is perfect (i.e., Pr(T)<1 for all but at most one method), then $EU(EMT)$ is strictly greater than $EU(DMT)$ as soon as there are 4 possible answers or more (3 or more if the number of methods is odd) and at least 3 methods. This is because in such cases there will always be possible outcomes where EMT will prescribe that the judgement should be withheld, while DMT will be more likely to endorse a false rather than a true result (see Appendix, Theorem A.1).[13]

The assumption $u(T) > u(W) > u(F)$ should be uncontroversial, as we explained above. Still, an objection could be raised that the result follows from our arbitrary decision to set $u(W) = 0$ and $u(T) = -u(F)$. As may easily be verified, the conclusion does not depend on these specific values. The only condition that needs to be satisified is: $u(W) \geq (u(T) + u(F))/2$. That is, as soon as the utility of withholding judgement is greater than or equal to the arithmetic mean of the utilities of endorsing a true or false result, EMT dominates DMT and thereby diffident purism. Again, it strictly dominates them when there are 4 possible answers or more (3 or more if the number of methods is odd) and at least 3 methods of which at most one is perfect.[14]

A straight-forward conclusion could be that our result relies on the idea that the value of withholding judgement should be exactly in-between the value of endorsing a false or a true result or closer to the value of endorsing a true result. A more critical reader could conclude that, of course, EMT will outperform DMT if withholding judgement is privileged in such a way. The above general result, however, only presents the lowest bound (a sufficient but not necessary condition) – $u(W)$ may typically be even closer to $u(F)$ than $u(T)$ and EMT will still have higher expected utility.

This is because when there are at least 3 methods and at most one is perfect (always providing true results) and there are 4 or more possible answers (3 or more for an odd number of methods), then DMT will be more likely to suggest a false than true answer (at least three-way ties or ties without the true result). The value of withholding belief can therefore, in all of these cases, also be closer to the value of endorsing a false than a true result and EMT will still come out as the more valuable approach. In fact, EMT dominates DMT as long as (for derivation see

[13] Note that we exclude the cases where there is only one method because one may only triangulate if there are two or more methods.

[14] It is worth mentioning that the same value of withholding belief, $u(W) \geq (u(T) + u(F))/2$, also comes up as the inflection point (i.e., qualitatively different strategies on either side) in Heesen and van der Kolk (2016), Appendix 2, in a different-context (a game-theoretic approach to peer disagreement). Although the example there does not directly relate to EMT, it presents interesting lines for future research regarding, e.g., gathering additional evidence in cases of discordant evidence (no unique result obtained by the plurality of methods).

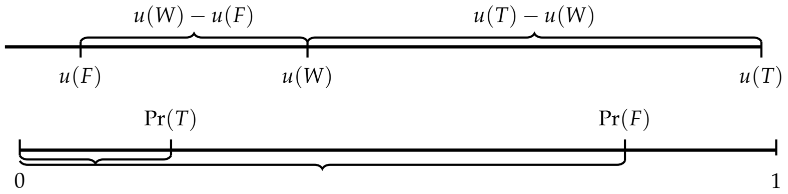

Figure 1: A representation of the inequality 11: The larger the ratio of $\Pr_{DMT}(F, \overline{dom})$ to $\Pr_{DMT}(T, \overline{dom})$ is, the closer $u(W)$ may be to $u(F)$ relative to $u(T)$ for EMT to still have higher expected utility than DMT

Appendix, Derivation A.2):

$$\frac{u(T) - u(W)}{u(W) - u(F)} \leq \frac{\Pr_{DMT}(F, \overline{dom})}{\Pr_{DMT}(T, \overline{dom})} \tag{11}$$

This means that the larger the probability of endorsing a false result due to methodological gambling is, the closer the value of epistemic modesty may be to the (in)value of falsehood (see Figure 1 for an illustration). Particularly, the probability of gambling upon a false result increases with a larger number of answers, which may be tied as this increases the number of potential methodological gambles with losing odds.[15] As a rule of thumb, the more possible answers there are, the less valuable epistemic modesty may be for EMT to still dominate both DMT and diffident purism.

How much is at stake when the judgement is withheld will depend on a value judgement of the triangulating scientist and especially on contextual determinants. However, as we have just established, even if the value of withholding judgement is closer to the (in)value of endorsing a false result than the value of endorsing a true result, EMT may still trump DMT and diffident purism. This provides a formal vindication of epistemic modesty in triangulation. Methodological gambling is not just intuitively implausible, it is also unacceptable given some plausible assumptions, e.g., that the value of withholding judgement is not too close to the value of endorsing a false belief.

6 Discussion

Our analysis rests on the assumption that it is more valuable to withhold belief than endorse a falsehood. This is because our focus is on epistemic qualities of EMT and not on practical decision-making where one may only decide to act or not to act. However, it may seem that our conclusion is already baked into the assumption that $u(F) < u(W)$ because methodological

[15] For example, suppose the correct answer is A. Then four methods with five possible answers (A, B, C, D, E) may provide answers: A, B, C, D. The probability of endorsing A by gambling in this outcome is only 1/4. If the answers are B, C, D, E, then it is impossible to endorse the correct answer through a gamble.

gambling implies that we will never be more likely to endorse a true rather than a false result in these cases. So what does our analysis add to the debate?[16] Moreover, what implications may withholding judgement have for practical evidence-based decisions?

It should be stressed that our formal analysis provides an insight in the specific value $u(W)$ on the continuum between $u(F)$ and $u(T)$. Similarly as Heesen and van der Kolk (2016, Appendix 2), we demonstrated that the arithmetic mean of $u(F)$ and $u(T)$ presents an important demarcation point: if $u(W)$ is greater than or equal to the mean of $u(F)$ and $u(T)$, then EMT always outperforms DMT. Our analysis, however, demonstrates even more. It shows that $u(W)$ may typically be closer to $u(F)$ than $u(T)$ for EMT to still come out as the better approach to triangulation. This is an interesting insight – even if withholding judgement is considered as a stance that is worse than neutral, EMT might still be preferable because it avoids methodological gambles with losing odds. This becomes especially clear if the methods operate with a larger number of possible answers because there will then be more potential situations in which the true answer is not even present.

By means of a toy example: Suppose 2 methods may provide 10 potential answers, of which only one is true. If we triangulate by DMT, then we will necessarily infer a true answer only in one potential outcome (when both methods provide the true answer). However, we will necessarily infer a false answer in the vast majority of possible outcomes where both methods provide false answers (in 81 of 100 potential outcomes). The remaining 18 outcomes depend on methodological gambling. The actual performance will obviously depend on the reliability of each method but because we assume that both DMT as well as EMT are to be used in states of methodological diffidence where the actual reliability of each method is not known to the scientist, the number of methods and answers presents a good rule of thumb in determining how low the value of $u(W)$ may be. Specifically, $u(W)$ may be closer to $u(F)$ with a decreasing number of methods and an increasing number of potential answers.[17] Hence, depending on how important it is to be opinionated (e.g., because of time-sensitive research) – which contextually determines $u(W)$ – and how many methods and potential answers there are, EMT may be preferable even if withholding judgement is considered to be an epistemic vice. That is, even if its value is closer to that of believing a falsehood than believing what is true.

By reducing the probability that a scientist would endorse a false result, compared to the use of a single method or DMT, our analysis also demonstrates that EMT serves an important role when it comes to improving epistemic trust in science. As Wilholt (2013) rightly points

[16] Thanks to an anonymous reviewer for bringing up this point.

[17] If the number of methods is larger, EMT becomes less competitive compared to DMT when the number of possible answers remains constant. Compare to the example of 2 methods with 10 possible answers another scenario where we have 20 methods and just 2 potential answers. In this case there will be only one potential case without the true result and fewer cases of methodological gambling because there are more methods than answers, so multiple methods will duplicate the same result.

out: epistemic trust is crucial for science and it depends on the reliability of the communicated results. The latter concept denotes the conditional probability that a result is true if a peer (scientist) declares it. It is easy to see that the conditional probability $\Pr(''A''|A)$ – communicating A when A is actually the case – is higher for EMT than DMT or DIFF. This is because EMT avoids communicating anything in what we called non-dominant outcomes (i.e., those where DMT randomly endorses one of the tied answers) where the probability of communicating a false answer is greater than or equal to communicating a true result.

It is now easy to see how the main difference between DMT and EMT spells out: the argument in favour of DMT revolves around its higher probability to provide a true result than methodologically diffident purism and, as we saw, also EMT. If our main goal is to increase the probability that we will uncover truth, then DMT is the way to go. The chief aim of science is, indeed, to find truth, but this is not all there is to it. Scientists need to also be appropriately geared toward truth, to borrow another Wilholt's term. If false results are communicated too often, then there is a genuine risk that the trust among scientists is broken. This is problematic for a number of reasons, efficient division of labour being but one of them.

The same also holds for public trust in science: if too many false results circulate, then this may have damaging real-world consequences once the falsehoods are uncovered. For instance, the public and the policy makers may become wary also of sound scientific advice due to unrelated past false results; a sentiment that is already too often abused (cf. climate change denial). We do not need to go further with examples to make clear why the goal of science is not just to provide true claims, but also to avoid falsehoods. What matters for our present purposes is that the questions related to epistemic trust in science provide a case in point of epistemically modest triangulation that we have proposed: when DMT provides some answer, it is more likely to be false than when EMT does. EMT is therefore more appropriately geared toward truth.[18]

Finally, is withholding judgement even an option that is always available? What implications might it have for scientific practice? Let us return to the example of determining someone's economic standing through social class, occupational status, individual earning, and total family income (Torche, 2011). Suppose that the answers are tied (2 or 4 different results by 4 indicators), which is a likely outcome because "distributions of these measures are only weakly correlated with each other" (Torche, 2011, p. 774). If we triangulate according to EMT, then we would discard these measurements and end with an impoverished data set. An objection might be made that our sample could become biased through this process because we avoid

[18] It is somewhat ironic that Du Bois, after whom DMT is named, also motivated his methodological prescriptions with what sort of procedures would allow scientists to avoid error and maintain public trust in science. It should be noted, however, that Du Bois assumed that a different concept is crucial for maintaining public trust, i.e., what he called pure-truth-seeking (see Bright, 2018).

inconclusive evidence.[19]

This is a reasonable worry: EMT favours caution over comprehensiveness, where by comprehensiveness we mean the ability to always provide a definitive answer. If evidence is too discordant, which is common in practice (cf. Torche, 2011; Stegenga, 2012), then EMT will err on the safe side and simply ignore it. This will often be a reasonable strategy: if evidence is unclear, then it is too unreliable. However, if too much evidence is discarded in this way, then its power will decrease. By means of our running example: a triangulating scientist might have many discordant data-points where different indicators point toward the individuals being of high and low economic standing, respectively. If there are too many such cases, the final reduced data-set may only contain the cases where it is clear whether individuals are of low or high standing. The sample might also become too small for the conclusions to have much statistical power when we correlate it with educational levels.

Moreover, we may worry that the cases of discordant evidence correlate with other variables of interest in the study, so ignoring them may lead to a biased sample (e.g., we might be excluding a specific category that for some reason produces discordant multi-method evaluations). A simple solution in such cases may be to gather additional evidence or use additional methods to break the deadlock. Unfortunately, this is not an option that is always available. It might be the case that additional evidence cannot be gathered (e.g., because the triangulating scientist is not able to collect additional evidence) or that the outcome depends on other pragmatic considerations (e.g., limited time and resources).

At least some of these worries are not new. Douven (2020), for instance, demonstrates that there is a trade-off between speed and accuracy when we assess the performance of different inferential rules. Some rules (an example of which is, broadly speaking, also EMT) are slower in recognising true hypotheses but more accurate, while others make bolder conclusions and thus lead toward truth faster but with a larger probability of embracing falsehoods. This may be an issue when it comes to time-sensitive inferences. An example of this (inspired by Douven, 2020) would be the inferential process of a doctor diagnosing a patient in an intensive care unit on the basis of multiple methods. Suppose evidence is inconclusive, but the doctor needs to figure out what is wrong with the patient. Our hypothetical doctor might run another round of tests, hoping to come to a clearer conclusion, but as time passes, the probability that the patient will not survive increases. It seems that in such cases DMT may be the preferred triangulation approach compared to EMT which might not be able to provide any diagnosis in time (due to, for instance, further evidence gathering).[20]

[19] Thanks to an anonymous reviewer for this observation.

[20] We leave further context-dependant explorations for future research which could be based on computer simulation simulations in a similar manner as those by Douven (2020).

All in all, this highlights that there might be situations in which withdrawing judgement will be too costly. Note that this is foreseen by our formal analysis: if $u(W)$ is too low, then EMT will not be preferred to DMT.[21] In this sense the two approaches might perhaps best be seen as complementary.

7 Conclusion

By introducing EMT we provided further reasons in favour of methodological triangulation. Moreover, we demonstrated that some of the pitfalls of its recently proposed version (DMT) may be avoided. This is important because methodological triangulation is a popular target for philosophers of science (e.g. Bovens and Hartmann, 2003; Claveau, 2013; Claveau and Grenier, 2019; Hudson, 2014; Stegenga, 2012).

We also need to stress that the advantages of EMT obtain even if all but one of the methods are only as good as mere guesswork. Although it is likely too pessimistic to conclude that most methods are just as reliable as pure guesswork, it is safe to say that there are not many methods (if any) that will always provide a true result. If there was such a method and a scientist was aware of its reliability, then obviously there would be no need to triangulate in the first place (moreover, the scientist would not be in a methodologically diffident situation). Crucially, however, EMT improves the outcome even if the methods are not very reliable. This is important because the methods used, for instance, in social sciences and psychology are also often unreliable. Bakker et al. (2012) estimate that a typical psychological study only has a power of about 0.35. In other words, if there is an effect, it will only be detected in approximately 35% of the cases. The fact that methods are often unreliable is also not alien to "hard sciences" (see, e.g., Elliott et al., 2020, for a recent example from neuroscience). Another reason why we should not overestimate the reliability of the methods is provided by the replicability crisis that is rampant in many scientific fields (including psychology, cancer research, and experimental economics; see Romero, 2019 for an overview). Hence, there are good reasons to consider triangulation in scientific practice because it can improve the performance of less-than-ideal methods.

As we demonstrate, however, it is also important to go with an appropriate way of conducting methodological triangulation. Specifically, EMT shows that epistemic modesty has a genuine place in scientific practice at least in some contexts. That is, unless there are good reasons to prefer comprehensiveness over avoidance of error, EMT presents a viable approach that benefits scientists who know that it is good to be in the right, but also risky to be in the wrong and are therefore willing to occasionally also withdraw judgement. After all, to know that you

[21] If $u(W)$ is too low to vindicate EMT over DMT, it will still be strictly greater than $u(F)$ because withholding judgement still prevents an outright error.

do not know is a sign of great wisdom as the wisest man in Athens already discovered many years ago.

References

Aaberge, R. and A. Brandolini (2015). Multidimensional poverty and inequality. In *Handbook of Income Distribution*, Volume 2 of *Handbook of Income Distribution*, pp. 141–216. Elsevier.

Bakker, M., A. van Dijk, and J. M. Wicherts (2012). The rules of the game called psychological science. *Perspectives on Psychological Science 7*(6), 543–554.

Borenstein, M., L. V. Hedges, J. P. Higgins, and H. R. Rothstein (2011). *Introduction to meta-analysis*. John Wiley & Sons.

Bovens, L. and S. Hartmann (2003). *Bayesian epistemology*. Clarendon Press.

Boyer-Kassem, T. (2019). Scientific expertise and risk aggregation. *Philosophy of Science 86*(1), 124–144.

Bright, L. K. (2018). Du Bois' democratic defence of the value free ideal. *Synthese 195*(5), 2227–2245.

Christensen, D. (2007). Epistemology of Disagreement: The Good News. *The Philosophical Review 116*(2), 187–217.

Claveau, F. (2013). The independence condition in the variety-of-evidence thesis. *Philosophy of Science 80*(1), 94–118.

Claveau, F. and O. Grenier (2019). The variety-of-evidence thesis: a Bayesian exploration of its surprising failures. *Synthese 196*(8), 3001–3028.

DiPrete, T. A. (2020). The impact of inequality on intergenerational mobility. *Annual Review of Sociology 46*.

Douven, I. (2020). The ecological rationality of explanatory reasoning. *Studies in History and Philosophy of Science Part A 79*, 1–14.

Du Bois, W. E. B. (1996 [1899]). *The Philadelphia Negro: A social study*. Philadelphia: University of Pennyslvania Press.

Elliott, M. L., A. R. Knodt, D. Ireland, M. L. Morris, R. Poulton, S. Ramrakha, M. L. Sison, T. E. Moffitt, A. Caspi, and A. R. Hariri (2020). What is the test-retest reliability of common task-functional MRI measures? New empirical evidence and a meta-analysis. *Psychological Science*, 0956797620916786.

Fleisher, W. (2018). Rational endorsement. *Philosophical Studies 175*(10), 2649–2675.

Fleisher, W. (2019). Endorsement and assertion. *Noûs*. In press.

Frances, B. and J. Matheson (2019). Disagreement. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2019 ed.). Metaphysics Research Lab, Stanford University.

Heesen, R., L. K. Bright, and A. Zucker (2019). Vindicating methodological triangulation. *Synthese 196*(8), 3067–3081.

Heesen, R. and P. van der Kolk (2016). A game-theoretic approach to peer disagreement. *Erkenntnis 81*(6), 1345–1368.

Hout, M. (1984). Status, autonomy, and training in occupational mobility. *American journal of sociology 89*(6), 1379–1409.

Hout, M. (1988). More universalism, less structural mobility: The American occupational structure in the 1980s. *American Journal of sociology 93*(6), 1358–1400.

Hudson, R. (2014). *Seeing things: The philosophy of reliable observation*. Oxford University Press.

Klein, D. and J. Sprenger (2015). Modelling individual expertise in group judgements. *Economics and Philosophy 31*(1), 3–25.

Koscholke, J. (2019). Robbers, pickpockets and average mutual firmness. *Analysis 80*(1), 45–51.

Martini, C. and J. Sprenger (2017). Opinion aggregation and individual expertise. In *Scientific Collaboration and Collective Knowledge*. New York: Oxford University Press.

Munafò, M. R. and G. D. Smith (2018). Repeating experiments is not enough. *Nature 553*(7689), 399–401.

Romero, F. (2019). Philosophy of science and the replicability crisis. *Philosophy Compass 14*(11), e12633.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin 86*(3), 638.

Schickore, J. and N. Hangel (2019). "It might be this, it should be that…" Uncertainty and doubt in day-to-day research practice. *European Journal for Philosophy of Science 9*(2), 31.

Skipper, M. (2020). Belief gambles in epistemic decision theory. *Philosophical Studies*, 1–20.

Stegenga, J. (2012). Rerum concordia discors: Robustness and discordant multimodal evidence. In L. Soler, E. Trizio, T. Nickles, and W. Wimsatt (Eds.), *Characterizing the Robustness of Science: After the Practice Turn in Philosophy of Science*, pp. 207–226. Dordrecht: Springer Netherlands.

Torche, F. (2011). Is a college degree still the great equalizer? intergenerational mobility across levels of schooling in the united states. *American journal of sociology 117*(3), 763–807.

Wiblin, R. (2020). How much does a single vote matter? *80,000 Hours*. https://80000hours.org/articles/how-much-does-a-vote-matter/.

Wilholt, T. (2013). Epistemic trust in science. *The British Journal for the Philosophy of Science 64*(2), 233–253.

A Appendix

Theorem A.1. *If A is the true answer and Pr(A)<1 for all methods except for at most 1, then if there are 4 or more possible answers (3 if the number of methods is odd) and at least 3 methods, DMT will more likely endorse a false than a true result when EMT will lead to withheld judgement.*

Proof. EMT prescribes that the judgement should be withheld if there is no unique answer obtained by the plurality of methods – we call this the non-dominant outcomes.

(i) Suppose there are three methods with three possible answers A, B, C. Then one possible outcome with a three-way tie is A, B, C. Because the result is endorsed by randomisation in case of a tie, the probability of obtaining a true compared to a false result is 1/3 to 2/3. EMT avoids endorsing any result. If $2n$ additional methods are added, there is always a possibility that n of them will provide the answer B, and n of them C, so the outcome is again non-dominant – now between the false B and C, so randomisation will necessarily lead to a false result in this possible outcome.

(ii) Suppose there are four methods with 4 possible answers A, B, C, and D. One non-dominant outcome is A, B, C, D. DMT will then only have 1/4 probability of uncovering truth, while EMT will suggest that no answer should be endorsed. This holds for any even number of methods greater than or equal to 4. If $2n$ more methods are added, there is always a possibility that n of the additional methods will provide answer B and n of them answer C.

Because (i) holds for 3 or more possible answers, it follows from (i) and (ii) that DMT will be more likely to endorse a false than a true result in some possible outcomes where EMT will lead to withheld judgement, given that there are 4 or more answers (3 if the number of methods is odd) and 3 or more methods. Hence, EMT will in these cases have strictly greater expected utility than DMT. □

Derivation A.2. *Derivation of the inequality 11:*

$$\frac{u(T) - u(W)}{u(W) - u(F)} \leq \frac{Pr_{DMT}(F, \overline{dom})}{Pr_{DMT}(T, \overline{dom})} \tag{11}$$

Proof. Let us assume that $EU(EMT) \geq EU(DMT)$, which we already proved for a special case (inequality 10). Hence:

$$\left(\Pr_{DMT}(T,\overline{dom}) + \Pr_{DMT}(F,\overline{dom})\right) u(W) \geq \Pr_{DMT}(T,\overline{dom})u(T) + \Pr_{DMT}(F,\overline{dom})u(F) \quad (12)$$

After some rearrangements we obtain:

$$\frac{\Pr_{DMT}(F,\overline{dom})\,(u(W)-u(F))}{\Pr_{DMT}(T,\overline{dom})} \geq u(T) - u(W) \quad (13)$$

Because $u(W) - u(F) > 0$ (Inequality 1) we complete the derivation with:

$$\frac{u(T)-u(W)}{u(W)-u(F)} \leq \frac{\Pr_{DMT}(F,\overline{dom})}{\Pr_{DMT}(T,\overline{dom})} \quad (14)$$

□