

Introduction

1. This annex is all about evidence assessment. It describes evidence assessment using the Evidence Framework (EF) a means by which evidence can be considered, evaluated and discussed. The EF comprises three simple tables, the Evidence Profile Table (EPT), the Validation Profile Table (VPT) and the Confidence Assessment Table (CAT); all three are produced at the end of this document. These simple tables form the Evidence Framework and they can be used for evidence assessment.

2. The EF is designed to help military desk officers, operational analysts, technical assurers etc assess and challenge the quality and validity of evidence used to inform decision. While the EF uses analytical language this is kept to a minimum, with statements being sufficient to help you understand the overall evidence quality required (the tables can be used in discussion with study stakeholders to set evidence targets) or the evidence quality achieved by a study.

3. While the thought of using three tables to help in assessment may seem to add unnecessary burden to busy workloads, the process, with practice, can be very quick and straightforward and is worth the investment as it enables conversations about evidence and assessments of "*how much evidence is enough?*" While this assessment is subjective it should not be seen as an inherent weakness. All evidence assessment is generally subjective in nature. What the EF provides through use of these tables is a means of challenging the evidence and a means of structuring the assessment of evidence, providing an effective audit trail to inform decision making.

4. It is possible to apply the EPT on its own as it says something about the quality of evidence and its assurance but it would not be sensible to use the VPT or CAT in isolation. The EF itself is consistent with analytical best practice; see the AQUA Book[1] for further details.

5. A worked example is provided below but before reading through the worked example it is worth taking note of some definitions. The EPT assessment results in something called a warrant which is used to say something about the quality assurance with respect to the evidence. **Warrant** is defined as "*evidence technical assurance using the categories weak, moderate, strong and proof/beyond reasonable doubt to describe the integrity of the evidence in relation to an assertion or hypothesis*". The assurance of the evidence is very much based on a study perspective, taking into account the limitations of the study, its scope etc. The VPT is about the validity of that evidence assurance in the wider context, i.e. taking into account the limitations with methods, measurement, what it is possible to know about the subject etc. Ideally an assessment of validity would be undertaken by people that have had little involvement in the study and can take a genuinely independent view from the outside looking in on the project, although it is recognised that this is not always possible. **Validity** is defined as "*a level of assurance that the right work is being or was engaged in using the categories weak, moderate, strong, high to describe the extent to which the work is fit for purpose in the wider context".*

[1] https://www.gov.uk/government/publications/the-aqua-book-guidance-on-producing-quality-analysis-for-government

The Evidence Profile Table (EPT)

6. The EPT is designed for use in assessing or evaluating the required or achieved quality of a body of evidence. The assessment is made in relation to an assertion or hypothesis. The assertion provides the relevant context and could be related to the fitness of the evidence to inform a particular decision, to evaluate the evidence derived from a methodology or method(s) to be used for a study or to evaluate the fitness of evidence from the study as a whole.

7. The EPT works by assigning a level between one and four to each of five evidence quality factors (considered to be generic characteristics of evidence), Comprehensiveness, Relevance, Challenge, Quantity and Veracity and taking the sum of these to provide an indication of warrant concerning the evidence. An EPT warrant is very much a study 'team' view, bound by the constraints on the project, constructed through sharing of the findings, methods used etc amongst peers to enable a judgement to be made about the quality of the evidence. The assessment supports transparency of evidence by the study team, but there is no reason why assessors external to the study team cannot apply the EPT to make their own judgements as part of a review process. The warrant can then be used to understand or assess the overall evidence position required or achieved for the assertion being made.

The Validation Profile Table (VPT)

8. The VPT is designed to complement the EPT. Its purpose is to assess or evaluate the validity of a body of evidence in relation to the assertion and is based on guidance contained within the AQUA Book[1]. While the EPT is essentially a study team view constrained by the boundaries of the study the VPT is best undertaken by assessors external to the study team, able to take a wider perspective on the validity of the findings. The VPT allows a judgement to be made regarding the extent to which the right work is being or has been engaged in, given the purpose and constraints placed upon that work. The key output from the validation process is a judgement concerning the extent to which the work is valid as part of the 'fitness-for-purpose' judgement.

9. The VPT works by assigning a level between one and four to each of four key validity criteria, Face Validity, Construct Validity, Content Validity and Criteria Validity and taking the sum of these to provide an indication of validity concerning the evidence. The validity score together with the warrant score derived through applying the EPT can then be used to understand or assess the overall evidence position required or achieved for the assertion being made. Both scores are used as indicators of 'fitness-for-purpose' and to estimate a position within the CAT to determine a confidence level.

The Confidence Assessment Table (CAT)

10. Whilst the EPT assessment will result in an evidence score and the VPT assessment a validity score there is often a need to express this in more simplistic terms and a need to understand the confidence in the findings. This is achieved by using the CAT to cross-reference the warrant inferred from the evidence score and the validity inferred from the validity score. Both are used to make a qualitative judgement about the confidence according to likely confidence bands. The confidence scale is "Very Low, Low, Medium,

High and Very High". Note, that the confidence shading is conceptual in nature to illustrate that boundaries are inherently fuzzy. In addition confidence should not be confused with probability ratings hence there is no quantitative expression of confidence.

11. The CAT works by taking the warrant score to position the assessment along the warrantability axis and the validity score to position the assessment along the validity axis of the CAT. There is a general rule of thumb associated with each warrant or validity criteria which provides a more informative statement about the judgement in relation to the findings. The intersection point is used to derive the confidence band.

A Worked Example

12. Consider a capability intervention where a decision is to be made regarding investment in a hypothetical new capability. For the purpose of the worked example the investment decision is supporting procurement assessment of the effectiveness of a new anti-tank guided weapon (SLINGER).

13. Prior to analysis being undertaken it would be possible to engage with relevant stakeholders to understand the target scores required for the evidence quality assurance expressed as warrant, the validity and the overall confidence. Undertaking an assessment at this stage would enable target levels to be set as part of the "*how much evidence is enough?*" conversation. These assessments can also be undertaken at the study planning stage, execution and exploitation stages to again assess evidence quality, validity and confidence.

14. So, regardless of the stage of the project how would an assessment be undertaken? Firstly, consider an appropriate assertion or hypothesis to test, which provides the context. For this example the assertion is that "*The introduction of SLINGER will improve the combat effectiveness of the battlegroup when compared to other possible alternatives for delivering effect*".

15. **Worked Example – The EPT Assessment:** The EPT factors are considered generic evidence characteristics and should be used as handrails around which to structure a conversation on evidence. Once each factor is scored simply sum the scores to arrive at a total that can be used to judge the extent of the warrant we can associate with the hypothesis regarding SLINGER. Let's now consider the factors in turn:

i. **Comprehensiveness**: For project SLINGER consider how many relevant factors can be taken into account given the constraints of time, resources, funding available to the project. For example, how many scenarios will be available for assessing SLINGER; how many different military actions can be assessed; how many of the system factors can be assessed to enable us to develop an understanding of the SLINGER system etc? Make a list of the factors considered relevant to judge the extent of the understanding that is attainable or that could be attainable. Once this is done choose the statements under this factor that best characterise the situation. For the purposes of the example assume this factor is scored as a **3**, i.e. we are not able to assess many of the relevant factors, due to SLINGER being at a relatively low technology readiness level and there may be some issues that could surface that are

not anticipated; the scenario set considered was limited due to time constraints and the target reaction represented by the combat models was claimed as representing behaviour that was appropriate but given the introduction of a SLINGER capability it is highly likely that the enemy would have changed its scheme of manoeuvre. So, for the SLINGER analysis the presentation of targets may have been over generous and hence may have provided more opportunities for engagement that may actually be the case. To improve the score consider increasing the coverage of the relevant factors, e.g. introduce more scenarios, explore more of the system level issues, consider course of action as an explicit factor to be explored or for more course of action variations to be analysed.

ii. **Relevance:** For project SLINGER consider how evidence drawn from a range of potential sources is relevant. For the hypothesis under consideration how relevant are previous studies, literature, data etc. It is also important to consider how the assumptions made would impact our understanding of the effectiveness of SLINGER and if when drawing conclusions about SLINGER how large is the inferential gap between assumptions and findings. So, for SLINGER consider carefully the context of sources, are there any inherent biases in the source material, are there appropriate perspectives that can help understand the system of systems view etc. Make a list of the key assumptions and assess the extent to which they may drive the findings, how would these assumptions affect the conclusions, would people have to make leaps of faith? Once this is done choose the statements under this factor that best characterise the situation. For the purpose of the example, assume this factor is scored as a **2**, i.e. many of the sources are relevant as anti-tank studies have been conducted in the past, historical analysis can be used to provide multiple other perspectives and although some of the assumptions have an impact it is considered that these are limited resulting in a small inferential gap from findings to conclusions. To improve the score would require some specific testing of the impact of varying assumptions to show the extent of the changes on the findings, wider consultation with a range of other groups to understand the impact of SLINGER more generally, e.g. impact on joint actions, environmental considerations etc.

iii. **Challenge:** For project SLINGER consider the extent to which findings have been or will be challenged and peer-reviewed prior to submission of the business case. This assessment helps to determine the extent to which the findings can be relied upon and how much challenge has been given to the findings based on the boundaries established by the project. Note that this is peer-review and challenge prior to wider socialisation of the findings with the wider stakeholder group. So, for SLINGER consider carefully the extent to which the findings are to be scrutinised, e.g. are the findings going to be scrutinised within project, more widely across the defence enterprise or through international fora to test the validity of the conclusions drawn? How are the assumptions and limitations recorded and caveated i.e. is there any appropriate data and assumptions paper produced for the project? Choose the statements under this factor that best characterise the situation. For the purpose of the example, assume this factor is scored as a **3**, i.e. the analysis is to be undertaken within Dstl who appoint a lead technical reviewer and the defence scrutiny

organisation appoints a lead scrutineer. However, review and scrutiny remains largely within the land domain resulting in review and scrutiny that is ultimately limited in scope. There is some recording of caveats and assumptions but no standalone data and assumptions paper has been produced covering all the major data items, hence there are some large limitations with respect to following the data audit trails. To improve the score consider widening the peer review circle to include those not directly related to the land domain to get a wider systems perspective, consider opening out the peer review to other nations drawing on the various exchange agreements to seek a range of comment. Also, consider commissioning a formal recording of data and assumptions in a form that is easy to read, challenge and check by relevant stakeholders.

iv. **Quantity:** For project SLINGER consider how balanced the methods for generating the evidence are. If there are a range of issues to explore then the project will likely need a variety of methods for generating the evidence, i.e. different qualitative and quantitative methods. Alternatively if this is not appropriate for project SLINGER, because there is a single method to be employed to determine effectiveness, then consider if it can be determined to be 'best-practice' through the extent of the track record for addressing problems of the type that SLINGER presents. It is not necessary to have a large quantity of sources to drawn on or multiple methods to score highly if there is a best practice method with a track record of use. Choose the statements under this factor that best characterise the situation. For the purpose of the example, assume this factor is scored as **2**, i.e. there will be a number of combat simulations run of varying levels of fidelity. These are considered best-practice for problems of this type and the model has a good track record for supporting combat effectiveness studies. In addition there will be judgement panels, looking at multi-criteria decision analysis[2] to generate other lines of enquiry. To improve the score consider the extent to which the lines of enquiry engage all the relevant stakeholders, i.e. it may be possible to use qualitative methods that are able to cover larger numbers of people and also methods that are better at exposing other issues, e.g. human factors operations, virtual mock-ups etc.

v. **Veracity:** For project SLINGER consider how consistent the evidence will be in relation to the wider evidential picture. Bearing in mind this is within the project boundaries, does the evidence form a highly supportive and integrated view? It is possible for there to be contradictions in some of the evidence and the extent to which these are explainable and can be integrated into a coherent evidential story is important. When results are generated for project SLINGER this considers the extent to which alternative accounts for the findings have been explored, i.e. if the battlegroup has improved effectiveness can we be sure that this is directly attributable to the introduction of SLINGER? It is important to consider the extent to which alternative accounts for the findings have been discounted. This plays directly

[2] **Multi-Criteria Decision Analysis (MCDA)** is a way of looking at complex problems that are characterised by any mixture of monetary and non-monetary objectives, of breaking the problem into more manageable pieces to allow data and judgements to be brought to bear on the pieces, and then of reassembling the pieces to present a coherent overall picture to decision makers.

to what can be said about cause and effect and if we can say that introducing SLINGER is directly related to improved effectiveness or not. Choose the statements under this factor that best characterise the situation. For the purpose of the example, assume this factor is scored as **3**, i.e. the results are somewhat consistent but it is difficult to weave this into a stronger integrated account of utility. Due to constraints on time and resources it was not possible to explore all alternative accounts for some of the findings and limited sensitivity analysis has been conducted. This means that for SLINGER we can only say that introducing SLINGER may cause the battlegroup to improve its effectiveness. To improve the score consider the wider narratives and if they raise any additional issues that could be explored as a part of a wider sensitivity analysis. Also, consider looking to improve the understanding of cause and effect by conducting additional analysis that increases the understanding of the system under study, i.e. explores some of the possible alternative accounts in more detail to understand if the effects seen can really be attributed to SLINGER.

16. **Assessing the warrant:** Now that all the evidence factors under the EPT have been considered simply sum the scores. So, for SLINGER the evidence profile is Comprehensiveness 3, Relevance 2, Challenge 3, Quantity 2 and Veracity 3. This gives a total of 13. Looking at slide 15 a score of 13 puts SLINGER firmly in the middle of Moderate warrant. When this is compared to the warrant axis on the CAT, slide 17, it shows that the rule of thumb for is that "Further evidence may change the findings". Taking the hypothesis that provided the context for the assessment we can say that the evidence quality assurance in support of testing the hypothesis "*The introduction of SLINGER will improve the combat effectiveness of the battlegroup when compared to other possible alternatives for delivering effect*" is considered to be of Moderate warrant (moderate assurance in effect) and that further evidence may change the findings.

17. **Worked Example – The VPT Assessment:** Now we understand the level of evidence assurance the project has produced using the warrant, given the project scope and limitations, we turn to assessing the validity of the findings in a more general sense, i.e. from outside the project looking in to understand the findings in a wider context. Understanding the wider context is important as it places SLINGER in the context of the wider system of systems. The VPT factors are considered generic validity characteristics and should be used as handrails around which to structure a conversation on validity. This conversation should really be one that is undertaken with external assessors who can bring a wider perspective, i.e. one beyond the boundaries of the project. Once each factor is scored simply sum the scores to arrive at a total that can be used to judge the extent of the validity we can associate with the hypothesis regarding SLINGER. Let's now consider the factors in turn:

i. **Face Validity:** For project SLINGER this considers the extent to which the findings and supporting arguments for the system under study are considered plausible. So, in essence has the analysis for project SLINGER passed the "do I believe it?" test for the recipient? Choose the statements under this factor that best characterise the situation. For the purpose of the example, assume this factor is scored as **2**, i.e. the findings and the supporting arguments are seen as largely

plausible and relatable to prior experience. To improve the score, would require consideration of the extent to which there are gaps in the arguments, e.g. is it arguments that are not clearly conveyed, are there some aspects that are difficult to relate to prior experience or for some reason are not seen as highly relevant and familiar to the recipients of the analysis because of the context of the study?

ii. **Criterion Validity:** For project SLINGER this considers how appropriate the inputs and outputs are for the system under study and also the extent to which the things being measured reflect the things being studied. So, for SLINGER to what extent can we be sure that the key input data aligned with that needed to represent SLINGER, the targets it would engage, the scenario etc. Also to what extent were the means of measuring the effectiveness of SLINGER or the other factors of interest valid? Did any of the combat simulations have too simple a representation of anti-tank engagements so that some of the purported measures of effectiveness can be called into question, e.g. some may only represent the chance of killing a target and not damaging a target, one model may not actually represent the details of engagement but has a proxy measurement for effectiveness based on rates of advance as result of combat power values etc. Choose the statements under this factor that best characterise the situation. For the purposes of the example, assume this factor is scored as **2**, i.e. the range of combat simulation models applied to the problem and the varying levels of fidelity have enabled the process of combat to be measured explicitly. Simulations are abstractions of the real world; however, the high fidelity simulation used for SLINGER assessment was able to measure performance of the system in terms of targeting, sighting, fly-out and engagement and how these contributed to engagement effectiveness. While not directly measuring actual variables of interest the surrogate variables are considered valid and show good alignment between the things being measured and the things being studied. To improve the score would require measuring actual variables of interest, so in the case of SLINGER this may involve setting up trials of early prototypes to measure the performance of the system with some of these measurements possibly being used to improve the combat simulations.

iii. **Construct Validity:** For project SLINGER this considers the degree of appropriateness of the key mechanisms that were used to represent the system under study, i.e. how good was the model 'construction' of SLINGER within the combat simulation (processes, relationships, structures). This is also about considering the extent to which model mechanisms are aligned to the current understanding of how SLINGER works. So, for SLINGER to what extent can we be sure that the modelling of any part of the engagement process has been represented to an adequate level and do the process relationships between the parts work as expected or are they adequate for purpose? Choose the statements under this factor that best characterise the situation. For the purpose of the example, assume this factor is scored as **2**, i.e. the model mechanisms while not explicitly representing SLINGER are largely appropriate, with the wider combat activities modelled relating as expected. For example, a number of relevant engagement sequences are represented and the relationships between the sequences are adequately represented and described. However, there are possibly

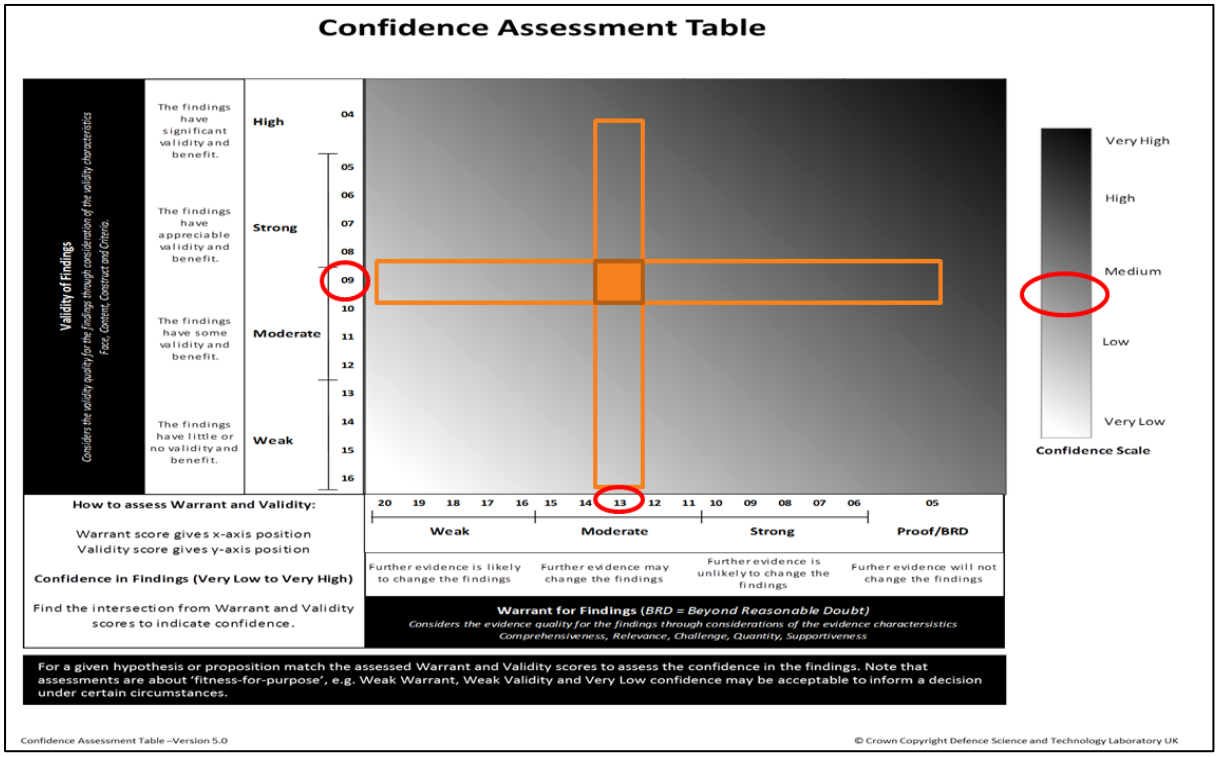
some key concepts and relationships not represented or not adequately represented which may affect the analysis. To improve the score would require higher fidelity representations of the key systems concepts and their relationships and an ability to fully describe these and the relationships. This may have to be through improvements to the models and their associated documentation to make it fit for the current purpose. Alternatively this may require new or bespoke methods developed to explore specific parts of the system to then provide data that can be used by the combat models.

iv. **Content Validity:** For project SLINGER this considers the extent to which it is possible to bridge the gap from findings to insight. So, for SLINGER testing the interpretative weight would ideally require engaging with communities outside of the immediate interests of the project boundaries to test if the findings are robust. Do they have sufficient fidelity in terms of their breadth and depth to support the insights? For example, this could be looking at Red Teaming the findings to look at effective countermeasures to SLINGER, taking into account wider integration issues etc. In addition this engagement would look to establish if the measurements were appropriately scaled and of sufficient granularity to warrant the claims regarding the findings, i.e. has there been over interpretation? Choose the statements under this factor that best characterise the situation. For the purposes of the example, assume this factor is scored as **3**, i.e. while the combat simulations have modelled relevant combat processes and have a sufficient set of variables to aid the measurement to generate some understanding what was claimed to be being measured is being questioned. For example, the targets presented during the combat operations modelling were being claimed as representing an appropriate set of targets for a SLINGER capability to engage. However, the modelling did not take into account wider air operations and it is highly likely that many of the targets presented for engagement by SLINGER may have been dealt with by air assets. So, for the SLINGER analysis the presentation of targets may have been over generous and hence may have provided more opportunities for engagement than may actually be the case. To improve this score would require reflecting on the assessment to see which additional factors would need to be assessed. So, in the example it would require target presentation rates to be an explicit factor to be explored or for more target variations to be analysed.

18. **Assessing the validity:** Now that all the validity factors under the VPT have been considered simply sum the scores. So, for SLINGER the validity profile is Face Validity 2, Criterion Validity 2, Construct Validity 2, and Content Validity 3. This gives a total of 9. Looking at slide 16 a score of 9 puts SLINGER firmly towards the top end of Moderate validity bordering on Strong validity. When this is compared to the validity axis on the CAT, slide 17, it shows that the rule of thumb for this is "The findings have some validity and benefit in relation to the problem under consideration". Taking the hypothesis that provided the context for the assessment we can say that the validity in support of testing the hypothesis "*The introduction of SLINGER will improve the combat effectiveness of the battlegroup when compared to other possible alternatives for delivering effect*" is considered to be of Moderate validity and that the findings have

some benefit in relation to the problem being considered but there are some limitations noted.

19. **Worked Example – The CAT Assessment:** The CAT assessment is relatively simple. Slide 17 shows there are two axes, the warrant axis and the validity axis. Take the warrant score of 13 and find the position on the warrant axis. Take the validity score of 9 and find the position on the validity axis. Look at the point of intersection and take the shaded area presented as an indication of the likely confidence we can have in the hypothesis "*The introduction of SLINGER will improve the combat effectiveness of the battlegroup when compared to other possible alternatives for delivering effect*". For the purposes of the worked example this would mean that we are tending to having medium confidence in SLINGER improving combat effectiveness, see the diagram below.



20. **Worked Example – Reporting the findings:** When reporting the result of the assessment for the hypothesis the following is suggested. There is **medium confidence** in the hypothesis "*The introduction of SLINGER will improve the combat effectiveness of the battlegroup when compared to other possible alternatives for delivering effect*". The assessment of confidence is determined by the quality of evidence assessment which indicates that evidence assurance is of **moderate warrant** and the assessment of validation which indicates **moderate validity**. Both ratings suggest that further analysis may change the findings but the current findings can be accepted as valid and beneficial.

21. The assessments using the three tables provide an appropriate audit trail for supporting the evidence claims for project SLINGER and can be considered analytical best-practice.

Evidence Profile Table

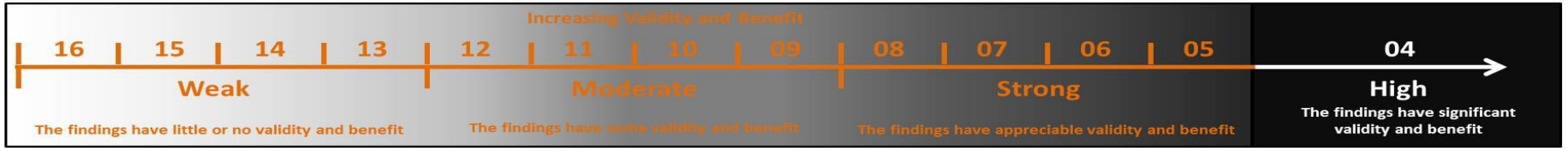
Comprehensiveness Considers the extent of the problem space that has or will be explored for the system under study as an indicator of the breadth and depth of coverage and understanding attainable.	Relevance *Considers the relevance of evidence (e.g. source studies, literature, data) and assumptions informing the findings for the problem currently being considered.*	Challenge *Considers the extent to which the body of evidence informing the findings has been peer-reviewed and independent challenge sought.*	Quantity *Considers the number and variety of sources as part of a balanced approach to the generation of evidence or the extent of the track record where variety is limited or unnecessary.*	Veracity *Considers the relation of the findings to the wider evidential picture, the extent to which alternative accounts for the findings are explored and what can be said about cause and effect.*	Profile Level
An extensive number of key aspects and related uncertainties have been or will be explored. System outputs and internal behaviour of the system can be described. An extensive number of the important processes in the system can be explained. Full or partial control of the system can be exercised under normal circumstances. Some system behaviour can be predicted or controlled under unusual conditions. This equates to a 'known knowns' perspective on the problem.	Evidence used to inform the findings draws from an extensive number of sources. These provide multiple relevant perspectives for understanding the wider context of the problem. Changes to the majority of relevant assumptions which could drive the findings have no impact on the utility of the findings for the current problem. There is assessed to be a very small inferential gap between evidence and findings for the current problem.	Review and scrutiny that has been or will be external to the study programme domain. This could be from across the wider department, organisation or from other relevant national or international organisations. These perspectives have been or will be able to provide extensive external challenge to the body of evidence used to inform the findings. Relevant caveats and assumptions have been or will be clearly stated. They do not or will not limit the utility of the work for its stated purpose.	The problem is complicated or complex, evidence has or will be drawn from a multi-method approach. This is through the extensive use of combinations of 'hard' and 'soft' methods. These provide multiple lines of enquiry to elicit multiple perspectives. Alternatively, the problem is well understood. Evidence has or will be drawn from a single or limited method approach. This is considered best practice with an extensive track record for addressing problems of this type.	Findings are highly related to the broader evidence base. Relevant evidence that has or will be taken into account forms a highly supportive and integrated view. Direct and indirect evidence used to support the findings is beyond all reasonable doubt. All relevant alternative accounts and views for the findings have been addressed and eliminated. In terms of cause and effect it is possible to say that the factor(s) A cause(s) the outcome(s) B.	1
The majority of the key aspects and related uncertainties have been or will be explored. System outputs and the majority of the internal behaviour of the system can be described. The majority of important processes in the system can be explained. Some changes in output or behaviour can be predicted for a limited time. Full or partial control of the system can be exercised under normal circumstances. This equates to a 'known unknowns' perspective on the problem.	Evidence used to inform the findings draws from a good number of sources. These have some relevant perspectives for understanding the wider context of the problem. Changes to the majority of relevant assumptions which could drive the findings have some but no significant impact on the utility of the findings for the current problem. There is assessed to be a small inferential gap between evidence and findings for the current problem.	Review and scrutiny has been or will be external to the study programme domain. This will be from across other relevant programme domains but not in the wider department. These perspectives have been or will be able to provide a good level of challenge to the body of evidence used to inform the findings. Relevant caveats and assumptions have been or will be clearly stated. To some extent they do or they will limit the utility of the work for its stated purpose.	The problem is complicated or complex, evidence has or will be drawn from a multi-method approach. This is through a good but limited use of combinations of 'hard' and 'soft' methods. These provide alternative lines of enquiry to elicit a variety of perspectives. Alternatively the problem is well or quite well understood. Evidence has or will be drawn from a single or limited method approach. This is considered good practice with a good track record for addressing problems of this type.	Findings are largely related to the broader evidence base. Relevant evidence that has or will be taken into account forms a largely supportive and integrated view. There is strong direct and indirect evidential support for the findings. Salient alternative accounts and some non-salient accounts and views for the findings have largely been addressed and eliminated. In terms of cause and effect it is possible to say that the factor(s) A is or are very likely to cause the outcome(s) B.	2
Some, but not the majority, of the key aspects and related uncertainties have been or will be explored. The nature of the problem space may be considered complex such that aspects are not easily explored. Some system outputs or some relationships between inputs and outputs can be described. Reliable prediction for a limited time is difficult. Reliable control is not possible. This equates to an 'unknown unknowns' perspective on the problem.	Evidence used to inform the findings draws from a limited number of sources. These provide a limited number of perspectives for exploring the wider context of the problem. Changes to some but not the majority of the relevant assumptions that could drive the findings have a significant impact on the utility of the findings for the current problem. There is assessed to be a large inferential gap between evidence and findings but it is asserted there is no doubt as to the value of their contribution for the current problem.	There has been or there will be some but limited challenge to the body of evidence used to inform the findings. Review and scrutiny has been or will be external to the study project team but within the programme domain. Relevant caveats and assumptions have been or will be clearly stated. These do or will largely limit the utility of the work for its stated purposes.	The problem is complicated or complex, evidence has or will be drawn from a single method approach. This is through a limited use of combinations of techniques within the set of 'hard' and/or 'soft' methods. This provides few alternative lines of enquiry reducing the variety of perspectives. Alternatively the problem is well or quite well understood. Evidence has or will be drawn from a single or limited method approach with a limited track record addressing problems of this type.	Findings are somewhat related to the broader evidential picture. Relevant evidence taken into account has been or will be used to form a somewhat supportive and integrated view. There is moderate direct and indirect evidential support for the findings. Most salient alternative accounts and some non-salient accounts and views for the findings have been addressed and eliminated. Some alternative accounts remain that could support the findings. In terms of cause and effect it is possible to say that the factor(s) A may well cause the outcome(s) B.	3
An extensive number of the key aspects and related uncertainties have not or will not be explored. The nature of the problem space may be chaotic meaning that aspects are difficult to explore or determine. It is very difficult to explain or predict system behaviour and control is not possible. Understanding is absent or very limited and equates to an 'unknowable unknowns' perspective on the problem.	Evidence used to inform the findings draws from a very limited number of sources. These provide a very limited number of perspectives for exploring the wider problem context. Changes to the majority of the relevant assumptions that could drive the findings have a significant impact on the utility of the findings for the current problem. There is assessed to be a very large inferential gap between evidence and findings and the current problem such that there is significant doubt as to the value of their contribution.	Review and scrutiny has been or will be within the study project team. As a result there has been or there will be very limited or little external challenge to the body of evidence used to inform the findings. Relevant caveats and assumptions have not been or will not be clearly stated. These do or will greatly limit the utility of the work for its stated purposes.	The problem is complicated or complex, evidence has or will be drawn from a very limited use of a technique within the set of 'hard' or 'soft' methods. This provides no alternative lines of enquiry and no variety of perspectives. Alternatively the problem is well or quite well understood. Evidence has or will be drawn from a single or limited method approach with no track record addressing problems of this type.	Findings show little or no relation to the broader evidential picture. Relevant evidence taken into account has not or cannot be used to form a supportive and integrated view. There is no or weak direct or indirect evidential support for the findings. Only some alternative accounts and views have been eliminated. Differently founded accounts are also assessed to have particular merit. In terms of cause and effect it is possible to say that the factor(s) A might cause the outcome(s) B.	4

For a given hypothesis or proposition consider each criteria in turn. Select a cell in each column that contains the statements that best describe the situation, noting that not all statements within a cell have to be relevant. Assign a score based on the Profile Level. Once complete add up the scores for each criteria. Compare the total score to the Warrant Scale to derive a Warrant statement expressing the degree of belief in the quality of evidence for the hypothesis or proposition.

Validation Profile Table

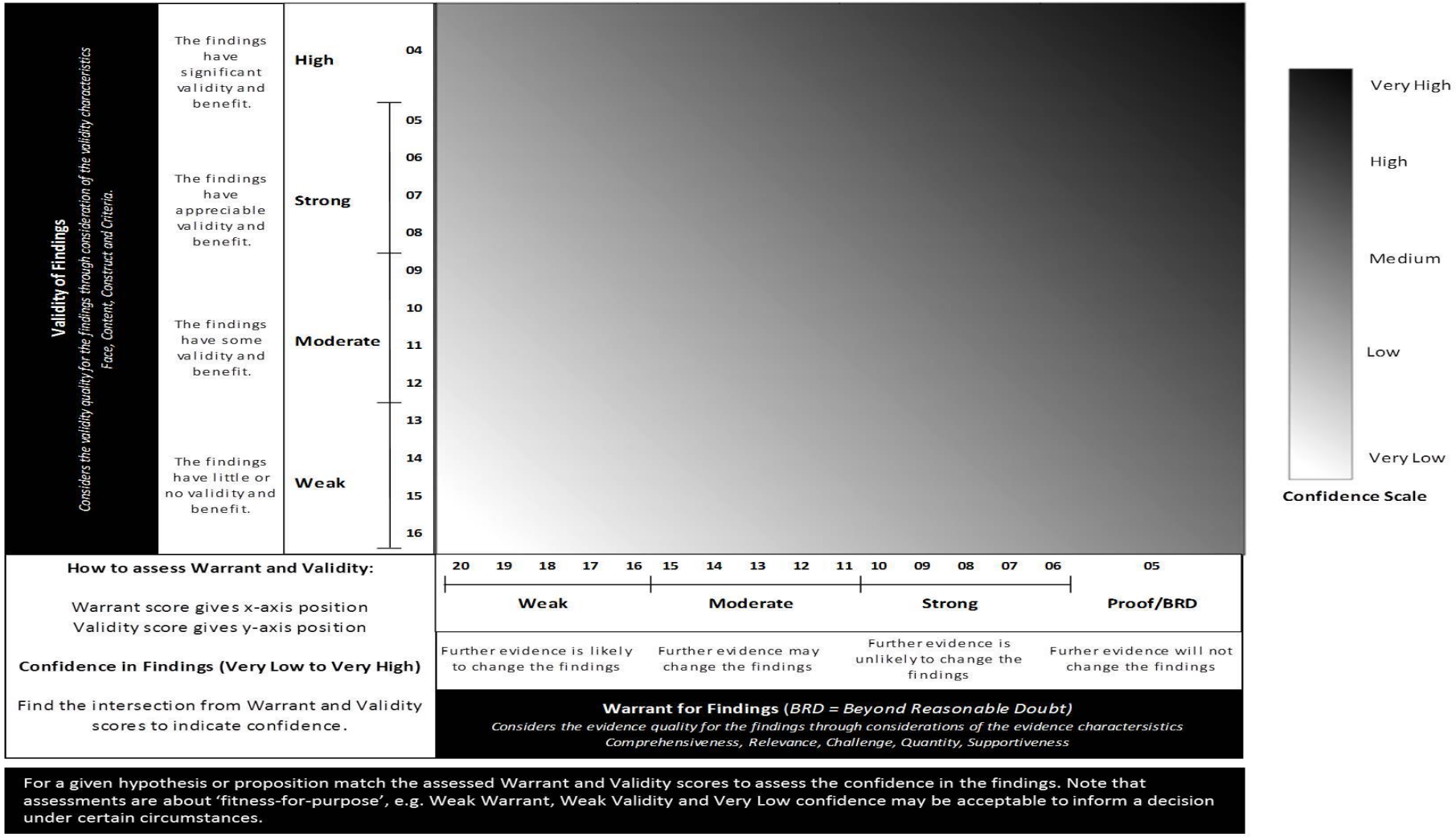
Face Considers the extent to which the findings and supporting arguments for the system under study are considered plausible. On the face of it do they pass the "do I believe it?" test for the recipient.	Criterion Considers how appropriate the inputs and outputs are for the system under study. Considers the extent to which the things being measured reflect the things being studied.	Construct Considers the degree of appropriateness of the key mechanisms. Those that will be or were used to represent the system under study.	Content Considers the extent to which it is possible to bridge the gap from findings to insight.	Profile Level
For the system under study the findings and the supporting arguments are seen or will be seen as highly plausible. They are seen or will be seen to be highly relevant and familiar to recipients of the analysis. They are seen or will be seen to be highly appropriate for the intended purpose. They are seen or will be seen to be highly relatable to prior experience.	The key input data used or that will be used for the system under study are well aligned with the data required to characterise the system. These are or will be highly suitable for the intended purpose. The analysis has used or will use actual system variables as outputs to measure the system under study. There is or there will be a strong alignment between the things being measured and the things being studied.	The key mechanisms used to represent the system under study are or will be highly appropriate. They are or will be highly adequate and sufficient for the purpose of addressing the problem. They are or will be strongly aligned to the current understanding of the system under study.	The findings have or will have high interpretive weight. This provides extensive insight into and a strong focus on the issues and the drivers relevant to the problem at hand. This is as a result of a high degree of fidelity in the findings, both in breadth and depth.	1
For the system under study the findings and the supporting arguments are seen or will be seen as largely plausible. They are seen or will be seen to be largely relevant and familiar to recipients of the analysis. They are seen or will be seen to be largely appropriate for the intended purpose. They are seen or will be seen to be largely relatable to prior experience.	The key input data used or that will be used for the system under study are largely aligned with the data required to characterise the system. These are or will be largely suitable for the intended purpose. Surrogate system variables have been used or will be used as outputs to measure the system under study. These are or will be largely suitable for the purpose of measuring the system under study. There is or there will be good alignment between the things being measured and the things being studied.	The key mechanisms used to represent the system under study are or will be largely appropriate. They are or will be largely adequate and sufficient for the purpose of addressing the problem. They are or will be largely aligned with what is currently known about the system under study.	The findings have or will have good interpretive weight. This provides good insight into and a good focus on the issues and the drivers relevant to the problem at hand. This is as a result of a good degree of fidelity in the findings, both in breadth and depth.	2
For the system under study the findings and the supporting arguments are seen or will be seen as somewhat plausible. There are or will be some limitations with the strength of arguments. They are seen or will be seen to be somewhat relevant and familiar to recipients of the analysis. They are seen or will be seen to be somewhat appropriate for the intended purpose. They are seen or will be seen to be somewhat relatable to prior experience.	The key input data used or that will be used for the system under study are somewhat aligned with the data required to characterise the system. These are or will be somewhat suitable for the intended purpose. The analysis has used or will use surrogate system variables as outputs to measure the system under study. These have been assessed as being somewhat adequate for the purpose of measuring the system under study. There is or there will be limited alignment between the things being measured and the things being studied.	The key mechanisms used to represent the system under study are or will be somewhat appropriate. There are or there will be some limitations with the structure of the mechanisms. They are or will be somewhat adequate and sufficient for the purpose of addressing the problem. They are or will be somewhat aligned with what is currently known about the system under study.	The findings have or will have some interpretive weight. This provides limited insight into and a limited focus on the issues and the drivers relevant to the problem at hand. This is as a result of a limited degree of fidelity in the findings, both in breadth and depth.	3
For the system under study the findings and the supporting arguments are seen or will be seen as largely implausible. They are seen or will be seen as largely irrelevant and unfamiliar to recipients of the analysis. They are seen or will be seen as largely inappropriate for the intended purpose. They are seen or will be seen as largely unrelated to prior experience.	The key input data used or that will be used for the system under study are largely not aligned with the data required to characterise the system. These are or will be largely unsuitable for the intended purpose. The analysis has used or will use surrogate system variables as outputs to measure the system under study. These have been assessed as being largely inadequate for the purpose of measuring the system under study. There is or there will be no recognised alignment between the things being measured and the things being studied.	The key mechanisms used to represent the system under study are or will be largely inappropriate. There are or there will be major limitations with the structure of the mechanisms. They are or will be largely inadequate and insufficient for the purpose of addressing the problem. They are or will largely lack alignment with what is currently known about the system under study.	The findings have or will have little or no interpretive weight. This provides little insight into and little focus on the issues and the drivers relevant to the problem at hand. This is as a result of little fidelity in the findings, both in breadth and depth.	4

For a given assertion consider each criteria in turn. Select a cell in each column that contains the statements that best describe the situation, noting that not all statements within a cell have to be relevant. Assign a score based on the Profile Level. Once complete add the scores for each criteria. Compare the total score to the Validity Scale to derive a Validity statement expressing the degree of belief in the validity of the hypothesis or proposition and the benefit in relation to the issue(s).



Validation Profile Table – Full Version 5.0

Confidence Assessment Table



Confidence Assessment Table –Version 5.0