FINDING METRICS FOR COMBAT AIRCRAFT MISSION EFFICIENCY: AN AHP-BASED APPROACH

J. Seethaler, M. Strohal, P. Stütz
Institute of Flight Systems, University of the Bundeswehr Munich, Neubiberg, Germany

Abstract

For assessment of multi-aircraft cooperation, e.g. in manned-unmanned teaming, total mission outcomes must be evaluated instead of simple single-aircraft parameters. In order to quantify mission efficiency (benefit and expense), a transparent and systematic metric derivation process is proposed. It involves the hierarchical decomposition of concrete missions by subject-matter experts (SMEs) to find measurable features. Subsequently, weightings are determined via pairwise comparison in a Fuzzy Analytic Hierarchy Process (AHP) using linguistic variables. An example study for one specific scenario was conducted as a proof-of-concept, which was evaluated positively by the SMEs.

Keywords
Mission Efficiency; Systems Effectiveness; Metrics; Analytic Hierarchy Process; Group aggregation

1. INTRODUCTION

Measures of performance for combat aircraft are needed to objectively assess and quantify their effectiveness and efficiency, i.e. benefit versus expense. This is required in aircraft development, for trade-off analyses [1], force deployment decisions, operational mission planning, requirements engineering, assessment and training of human operators, and as utility functions for machine learning. Such assessments play a major role for governmental decision makers in comparing different aerial platforms to find the one best suited for a set of missions when deciding on the procurement of a successor to a current (manned) aircraft.

However, in today's complex mission environments, requirements for performance indicators go beyond simple single-aircraft parameters and properties [2]. Furthermore, massive interdependence is introduced when evaluating composite forces of multiple aircraft – such as wingman configurations, combinations of unmanned aerial vehicles (UAV), manned-unmanned-teaming (MUM-T), or fully autonomous swarms –, because significant interaction occurs. Here, a composite is made up of cooperating distributed subsystems, i.e. the individual aircraft, and is called a *system of systems*. An example for said interaction is communication between entities, when sensing and effecting capabilities are separated ("sensor-to-shooter") and e.g. jamming will adversely affect the mission performance. Especially of interest for distributed systems is how the benefit (e.g. by the expected higher situational awareness [3]) of additional entities scales with their number.

Consequently, for finding out how well a certain system-of-systems performs at fulfilling its purpose, the execution of missions, more precisely the mission results, must be judged. In that way, *measures of performance* (MoP) of individual systems are superseded by overall *measures of effectiveness* (MoE) for the cooperating entities. Then, by quantitatively assessing all relevant missions including variations of technical and non-technical parameters, e.g. plans or opposing forces' actions, an overall assessment of each considered system-of-systems can be achieved. This enables comparing the quality of alternative platforms for the same set of missions.

That gives rise to the need to quantify benefit and expense for concrete missions – either real or simulated – first. The quantification methodology must be agnostic to number and type of aircraft and equipment; it needs to be applicable to manned and unmanned aircraft alike, in both homogeneous and heterogenous force packages.

This work presents an approach to find useable and relevant metrics for a multi-aircraft mission in a structured, systematic, and approachable manner. The proposed method makes extensive use of subject-matter experts' (SMEs) knowledge. It also avoids creating a "black box" by being transparent. A proof-of-concept study for one specific scenario has been conducted and evaluated. The need for a mission evaluation framework aimed at force packages and composites of multiple aircraft is also highlighted by the SME evaluation of the proposed method in chapter 6.3.

2. STATE OF THE ART

Assessing a weapon system's effectiveness and efficiency is part of the long-standing domain of military *operations research* (OR) [4], where metrics serve as quantitative indicators for the quality of a system. In OR the rather general concept of *Systems Effectiveness* [5], meaning the probability of success under given circumstances, has been heavily used. It is based on the so-called "-ilities", mainly availability, capability and reliability, which directly and indirectly affect the outcome of a mission and depend on a systems' technical properties and human factors. Assessments can range from single entity scale, e.g. a single aircraft, up to a joint (ground, air and maritime) campaign scale on theater level [6].

In OR, modelling and simulation often is used as sole or additional data source. To that end extensive simulation environments have been set up and widely used in the armed forces and by industry, e.g. the multi-agent based environment PAXSEM by Airbus [7]. This simulation is used for stratified Monte Carlo sampling under the headline "data farming" [8], where key performance indicators (KPIs) are measured, weighted and aggregated, but not transparently nor systematically derived.

Additionally, in the domain of artificial intelligence (AI) and machine learning, utility functions and measurements of an AI's success are required in reinforcement learning or genetic algorithms [9]. However, these are often only geared towards driving the AI in the desired direction and

doi:10.25967/530013

thus are without any real world meaning or impact on human decision makers when evaluating a system's effectiveness e.g. in a procurement decision.

For such use cases the University of the Bundeswehr Munich (UniBwM) Institute of Flight Systems (IFS) researches on aircraft assessment techniques on various scales. It has previously developed concepts and more specifically multi-criteria methods for single aircraft comparisons, which also refer to specific missions, but are focusing on single-aircraft parameters and properties rather than on explicit constructive simulation of the scenario [10, 11]. However, the latter becomes a necessity in multi-aircraft composites because of the heavy interactions and nonlinear interdependencies of cooperating and communicating systems.

Thus, this work is concerned with assessing combined forces consisting of cooperative entities, also by using multi-criteria methods, but specifically focusing on mission outcomes based on the systems effectiveness idea. We want to present a methodical approach to define what measurable/observable success in a combat aircraft mission means. It shall be more structured and transparent than previous assessment methods regarding distributed systems and enable insights mainly for mission planning, in procurement and systems development.

3. FUNDAMENTALS

As stated before, an evaluation framework for composites of aircraft needs to be agnostic to number and type of aircraft (e.g. manned or unmanned). Also, the consequences of interactions in the system of systems need to be captured. Furthermore, the composite usually interacts with other blue and red forces during a mission. Therefore, only mission outcomes, i.e. results, should be considered, but not the way these outcomes were achieved.

With the help of subject-matter experts (SMEs) and their a-priori knowledge a specific mission can be structured and decomposed into its elements. Thus, the missions' desired and undesired outcomes can be identified. Measuring efficiency, i.e. benefit in relation to expense, also demands the criteria of cost to be defined.

To aggregate the criteria to a single, unified mission effectiveness/efficiency rating their respective interrelations must be considered. Mainly, their relative importance has to be determined by linking all criteria on each level of decomposition with weightings. Thus, having gained a rule on how to combine the previously found criteria, a mission's success can be measured by measuring all its elementary criteria.

When judging an aircraft or composite of aircraft multiple possible ways of executing a mission should be averaged accounting for possible variations in allied, neutral, and/or enemy influence. Because there are different possible approaches or plans for carrying out the mission, several must be tried and averaged, if one wants to assess the technical characteristics of a system. Ideally, only the optimal plan would be considered for this purpose Finding this optimum also requires a quality measure, which again is the aggregated mission effectiveness or efficiency metric. A possible optimization strategy is to use *Pareto fronts* [12].

In the end, all missions must be weighted themselves and aggregated to gain a system of systems' total assessment. When the same missions are used throughout, direct comparisons of entirely different systems of systems are possible.

4. CONCEPT

Figure 1 illustrates the basic concept on how to gain a mission-specific assessment comprised of three main blocks: First stands the *SME metric derivation*, on which this work focuses on. The experts decompose the objectives of a given mission (as well as effort items) hierarchically in a *tree structure* with nodes and sub-nodes. Eventually sub-goals and cost elements on a certain elementary level cannot be further decomposed but are easily measurable observables. Then, the stakeholders or SMEs can order and thereby weight these sub-goals at each node of decomposition, using simple linguistic variables in a Fuzzy Analytic Hierarchy Process [13] (FAHP). Applying such Multi-Criteria Decision Analysis (MCDA) method, the assessment of the mission's outcome (also known as *end state*) becomes a well-defined problem. The systematic, hierarchical breakdown of the specific mission supports comparability and maximizes objectivity, while the weighting process also allows for decision makers' preferences to be included transparently and traceably.

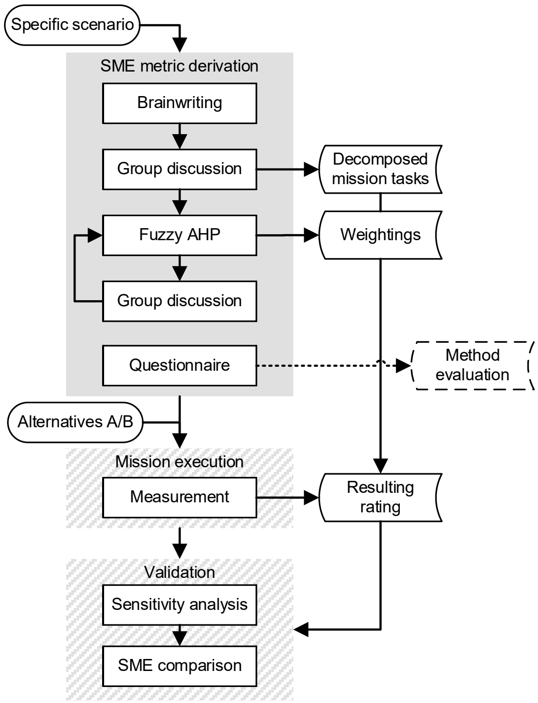

Figure 1: Process flow diagram of systematic metric derivation, simulation, and validation.

The metric derivation yields a total objective function (the *quality integral*) as a measure of the system's effectiveness. Its general form is given in equation 1:

$$J = \sum_{k=1}^{N} \sigma_k w_k j_k \qquad (1)$$

The overall effectiveness J is a sum for all $k=1,\dots,N$ elementary criteria over their respective (global) weight w, the elementary criteria evaluation value j (normalized to be in the interval $[0,1]$), and the respective sign $\sigma=-1,+1$, which is positive one for benefit ("good") and negative one for cost and effort ("bad"). Efficiency is defined as effectiveness divided by cost, i.e. the total J divided by the absolute value of the sum of all elements with negative sign σ.

The second block is the execution of the mission. Once the metric for the specific scenario is derived, i.e. the mission decomposed and its components and tasks weighted, the observables can be measured in a simulation or even the recording of a real-life mission. Specific measurement routines need to be implemented. The subsequent aggregation results in the mission-specific rating of the considered system of systems. This quantitative rating can be used to rank and compare several alternative systems of systems.

Finally, in the third major block of *validation*, the previously derived metric needs to be checked. The course of the mission should be presented to the SMEs from the original metric derivation process along with the resulting rating. Local and possibly global sensitivity analysis of resulting and intermediate ratings and the weightings should be conducted to show how strongly the total resulting rating is influenced by deviances in the weightings per decomposition node.

According to the focus of this paper, the following paragraphs now describe the systematic metric derivation in more detail.

4.1. Brainwriting and Group Discussion

Before the metric derivation starts, the concrete mission must be well defined. The specific scenario should be broad enough to allow for some variation, but also narrow enough to provide the SMEs with a relatable, realistic mental image. Initially, all SMEs are briefed about the scenario either at the same time or at least with exactly the same information. Exact numbers (e.g. spatial distances or numbers of potential adversaries) do not need to be provided to the SMEs, but a general idea of the strategic and tactical situation is required.

Subsequently the mission structure must be established. Here the SMEs decompose the mission hierarchically in its objectives and sub-objectives, until measurable elementary criteria are reached. A tree structure is a helpful visualization. First, in a short session of brainwriting each individual expert is encouraged to decompose the mission independently from the other SMEs. *Brainwriting* [14] means every participant takes notes separately, as not to be not influenced by others. The idea is to avoid suppression of opinions or knowledge by group dynamics, and to save information for later by writing it down.

The following step is about finding a consensus in an open *group discussion*, referring to the notes from the brainwriting step. The objective is to define one unified hierarchical decomposition of mission criteria. The final derived structure needs to be the same for all SMEs, otherwise comparability and compatibility is lost. At this stage, moderation by an independent person is advisable. Terms and phrasings need to be agreed on in the group, otherwise imprecisions ensue. Use of graphical support by e.g. a mind-mapping software or plotting a tree structure on a whiteboard or a flipchart proved to be helpful. An example result is shown in chapter 6.1.

In this stage the experts also need to determine the *optimization targets* for each criterion, i.e. to determine whether the value of its measurement needs to be minimized (effort or cost criterion, $\sigma=-1$), maximized (benefit criterion, $\sigma=+1$), satisfied (must be smaller/larger than a certain threshold), or fixated (only one exact value is acceptable). Finally, the normalizations and the shapes of the utility functions need to be determined. However, it might be advisable to revisit these in the group discussion after the weighting stage.

4.2. Weightings

Subsequently, the weights w_i of all mission elements are to be established. These weights symbolize the importance of the mission criteria and should be normalized for each decomposition node. The measured values of the lowest-level, i.e. elementary, criteria are multiplied with the weight and then summed up, all children nodes yielding a parent node's value. Adding values along all levels until the whole mission (i.e. the top node) is reached gives the final numerical value of mission effectiveness. Naturally, negatively perceived items such as cost receive a negative sign.

To quantitatively determine the weights, each SME uses the *Analytic Hierarchy Process* (AHP) weighing all the children of one parent node against each other, over all decomposition nodes and levels.

4.2.1. The Analytic Hierarchy Process (AHP)

The *Analytic Hierarchy Process* [15] is a compensatory method from decision making theory. The idea is to break the decision criteria down into a hierarchy and then to derive relative preference weights for all criteria by pairwise comparison [16].

In such pairwise comparison, the dominance of one criterion over the other is expressed on a numerical scale, usually but not necessarily using integers ranging from 1 to 9 (and their reciprocals). The comparisons of all n items on one node make up a square $n\ x\ n$ comparison matrix A, each element a_{ij} representing the importance of criterion i over criterion j. Naturally, all elements are self-equivalent ($a_{ii}=1$), meaning the diagonal is always unity. Off-diagonal elements (with row and column indices i and j for criterion i versus j, with n criterions total) are always automatically reciprocal:

$$a_{ij} = \frac{1}{a_{ji}} \quad \forall i, j \in \{1, \ldots, n\} \tag{2}$$

As an example: if criterion 1 is deemed much more important than criterion 3, one could assign $a_{1,3} = 7$, automatically yielding the reciprocal matrix element $a_{3,1} = 1/7$. From that matrix of pairwise relative importance rankings, the overall local weights w_i of the criteria are calculated by solving for the principal eigenvector $(w_1, \ldots, w_n)$ and then normalizing, as per equation 3 [15]:

$$A\vec{w} = \lambda_{\max}\vec{w}, \qquad \vec{w} = \left(w_1, \ldots, w_n\right) \tag{3}$$

In equation 3 $\lambda_{\max}$ is the principal eigenvalue, i.e. the largest of all eigenvalues* λ_i, $i=1,\ldots,n$. Methods to derive the weights other than the eigenvalue method are not recommended [18]. The global weight of a decomposition node is then gained by multiplication of its local weight with all its parent nodes' local weights.

As subjectivity plays a role in the pairwise comparisons by each individual SME, the c*onsistency index* (CI) [15] serves as a measure of consistency:

* One finds the eigenvalues of a matrix via solving the characteristic polynomial equation $det(A-\lambda E)=0$ [17].

(4) $$CI = \frac{\lambda_{\max} - n}{n-1}$$

Furthermore, the *consistency ratio* (CR) is given by equation 5, where RI is the average consistency index of randomly generated comparison matrices [16]:

(5) $$CR = \frac{CI}{RI}$$

Values of RI for comparisons of up to ten items are detailed in the appendix. The matrix A is fully consistent if transitivity $a_{ij}a_{jk}=a_{ik}$ holds for all indices i,j,k, giving $CI=0$. If CR exceeds 0.1, the comparison matrix should be reexamined, so in that case an iteration loop with the respective SME is recommended. Sometimes a CR up to 0.2 is seen as still acceptable [19].

4.2.2. Fuzzy AHP

Due to inherent imprecisions in human decision making and judgement, it can be hard for SMEs to attach a numerical value to their strength of preference for a criterion. Also, slight individual differences and nonlinearities in perception may occur. To better depict this in the ordinal paired comparison, we recommend using linguistic variables instead of fixed numerical values. Taking the idea from fuzzy logic, in *Fuzzy AHP* (FAHP) these comparative linguistic variables are projected onto a numerical scale via partially overlapping fuzzy sets [13].

A scale of seven linguistic variables and their respective triangular fuzzy set mapping to a numerical scale from 1 to 9 are shown in table 1.

Linguistic variable	**Triangular fuzzy set** (l,m,u)
Absolutely dominant	(8, 9, 9)
Very much more	(6.5, 7.5, 8.5)
Much more	(5, 6, 7)
Substantially more	(4, 5, 6)
Significantly more	(3, 4, 5)
Slightly more	(1.5, 2.5, 3.5)
Equivalent	(1, 1, 2)

Table 1: Linguistic variables for ordinal paired comparison and their respective triangular fuzzy sets for Fuzzy AHP, in accordance to [13].

Such triangular fuzzy set is defined by a tuple of three values (l,m,u) according to the membership function given in equation 6:

(6) $$\mu(x) = \begin{cases} \dfrac{x-l}{m-l}, & l \le x \le m \\ \dfrac{u-x}{u-m}, & m \le x \le u \\ 0, & else \end{cases}$$

To find the weights via the eigenvalue method shown in the previous chapter, each fuzzy comparison matrix with tuples $(l,m,u)_{ij}$ as elements is split into three non-fuzzy (i.e. sharp) comparison matrices with elements l_{ij}, m_{ij}, and u_{ij}, respectively. The three principal eigenvectors (each of length n for n criteria) combined give n triangular fuzzy sets, which are the fuzzy weights $(w_l,w_m,w_u)_i{\equiv}(w_{il},w_{im},w_{iu})$. Then, to gain a usable, sharp local weight value w_i for each criterion, so-called *defuzzification* is necessary. A possible method is given in equation 7 [13]:

(7) $$defuzz(l,m,u) = \frac{\frac{1}{3}\left(-lm + um - l^2 + u^2\right)}{u-l}$$

One has to ensure that $w_l \le w_m \le w_u$, which can lead to additional steps [20]. Some authors also suggest different procedures in FAHP, e.g. [21]. FAHP also allows for consistency checks as seen in chapter 4.2.1 and requires it as much as standard AHP.

4.2.3. Robust Simos method

To even better represent the subjective expertise and personal preference of the respective SME, the individual weighting results can be manually modified after the ranks are determined by FAHP. For this purpose the modified *Robust Simos method* [22] for adapting the weight distances in between the ranked items was implemented. However, in the experiment only one SME chose to use this method and subsequently dismissed it as not having additional benefit over pure FAHP.

4.2.4. Iteration

As illustrated in figure 1, another group discussion follows the FAHP. The feedback loop to the FAHP step ensures better consensus. All SMEs are presented the resulting weights, their respective spread, and consistency ratios graphically, e.g. via box plots. Relatively large spreads in the weights are signs of disagreement, either in opinion about the priority of a (sub)criterion or in its definition. Open discussion can yield further decomposition of the respective node, an updated definition of the criterion's meaning or naming, and/or changes in the FAHP comparison matrix and thus the weighting by single or multiple SMEs. Even alterations in the decomposed mission structure might be required, as further interdependencies can be discovered in this later stage.

This iteration loop of returning to FAHP and then discussing the results can be repeated as often as needed. Especially when single or multiple SMEs have a CR higher than 0.1 in one or more nodes, the specific comparison matrix should be carefully (without changing preference rank order) modified accordingly.

4.2.5. Weight Aggregation

Multiple group aggregation techniques for AHP have been suggested in literature concerning group decision theory.

One approach is the *aggregation of individual judgments* (AIJ) [23], where the preference values a_{ij} in the comparison matrix are aggregated either via arithmetic or preferably [24] the geometric mean. Only then the preference order and weights are computed via the eigenvalue method.

Conversely, in the *aggregation of individual priorities* (AIP) method [23] one first calculates the overall weights w_{ir} from the comparison matrix as shown in 4.2.1 for each expert r. From those the aggregated weight value $\overline{w}_i$ for criterion i is calculated then via the weighted arithmetic mean, equation 8, or the weighted geometric mean, equation 9.

(8) $$\overline{w}_i = \frac{1}{c}\sum_{r=1}^{R} w_{ir}\, p_r$$

(9) $$\overline{w}_i = \frac{1}{c}\prod_{r=1}^{R} w_{ir}^{\;p_r}$$

The normalization is symbolized as division by the constant c. In both AIJ and AIP the importance of individual SMEs can optionally be included by weighting their contribution to the mean by a factor p_r for a total of $r=1,...,R$ SMEs.

AIJ sports several weaknesses in respect to relevant rationality axioms compared to AIP [23]. Also, AIP can rather easily be used with the additional *Simos* modification and more importantly is not computationally intensive if FAHP results of additional SMEs should be added at a later time.

Lastly, the *loss function approach* (LFA) [23, 25] explicitly takes the measure of consistency into account, by relating each SME's consistency ratio CR to the mean and the variance of all CR and thereby prioritizing more consistent SMEs. LFA is more complicated to implement than AIJ and AIP. Also, with comparison matrices for up to two criterions CR is not well defined, as $RI=0$ for $n=1,2$, leading to obvious issues in calculating the mean and variance of the CRs.

5. METHODICAL IMPLEMENTATION

As a proof-of-concept the method presented in chapter 4 and visualized in figure 1 was implemented and tested with a small set of five SMEs ($R=5$). The SMEs had a mean professional experience of 5.3 years in the area of aircraft assessment and 2.8 years in mission result analysis. Additionally, some SMEs had experience as aircraft operators and others in military anti-air defense.

5.1. Demo Scenario

As a demo scenario a relatively simple mission vignette was chosen, describing a basic air-to-ground operation. In a given zone of operation in enemy territory a convoy of vehicles, e.g. a terrorist/insurgent threat, needs to be found and neutralized. In the operational zone there are pop-up threats against the aircraft by surface-to-air missiles (SAM) including *Man Portable Air Defence Systems* (MANPADS) of unknown location and number. Opponent fighter aircraft are assumed to be on standby nearby. Possible neutral and civil entities must be respected. Covertness, especially during ingress, is desired. The number and type of own aircraft was unspecified, in order to focus on mission outcomes and not on aircraft properties. Own airships in this scenario will interact mainly by communicating with each other about their findings.

Because the vignette is rather unspecific, it was not only described to the SMEs, but also discussed in the group to further define the constraints and boundary conditions prior to commencement of the metric derivation process.

5.2. Computer aided Fuzzy AHP

A computer program was developed with an easy-to-use graphical user interface (GUI) for FAHP and the Simos method. The only necessary input to start is the mission tree, i.e. the decomposed structure. In this GUI, one decomposition node is active at a time. Its comparison matrix is shown in full, along with all the available linguistic variables from table 1, chapter 4.2.2. Once a comparison is made via selecting a linguistic variable, the software automatically fills in the appropriate reciprocal value. When the comparison matrix is complete, the program immediately calculates the weights and displays the consistency index CI and ratio CR. The ranks and weights are displayed graphically and numerically. Optionally, the distances can be modified via drag-and-drop according to the Simos method (chapter 4.2.3).

After all SMEs have completed all comparison matrices independently, the software can generate boxplots (minimum, maximum, median, first and third quartiles) of the local weights for all (sub-)criteria and CR at all nodes. The aggregation techniques AIJ, AIP and LFA were implemented. AIP was chosen as most suitable, as described in chapter 6.2.

5.3. Expert Questionnaire

#	Statement	Answers favorable +/o/-
1	The explicit assessment of aircraft mission results is very relevant to me.	+
2	My previous approach is more based on intuition and experience than hard rules/methods.	o
3	My previous approach for assessing mission results is well structured.	o
4	The results of my previous approach are unrestrictedly comprehensible/traceable.	+
5	My previous approach is fully adequate for assessing missions of groups/composites of aircraft.	+
6	For assessing groups/composites, assessing all aircraft only individually is adequate.	+
7	The previous approach significantly takes into account interaction between aircraft in the group.	+

Table 2: General situation and previous approaches: Statements presented to SMEs and favorability of their answers towards the demonstrated new method.

Outside of the actual proposed method, but also represented in figure 1, the SMEs were presented a questionnaire for evaluation of said method. The first set of questions concerning the relevance of the issue and their previous approach to judging mission outcomes and composites of multiple aircraft (see table 2 and figure 2) was to be answered before the metric derivation. The second set of questions (see table 3) was presented after the metric derivation process and concerned the evaluation of the method regarding its structure, completeness, preferability versus the previous approach and effort. Both sets of were given as statements and a five-point *Likert scale* [26] each.

6. RESULTS ON DEMO SCENARIO

Figure 2 shows the answers to the first set of questions. Table 2 also denotes (symbolized by +/o/-) whether the SME's answers were favorable towards the approach in this work. The questionnaire revealed that while the SMEs largely have a need for explicit assessment of aircraft mission results, they currently did not have a well-structured, fully traceable (transparent) approach. In fact, they partially rely more on experience and intuition when judging aircraft mission outcomes. Even more importantly, there was significant agreement that only assessing single aircraft individually is not adequate for judging systems-of-systems-style composites. The SMEs thought their previous approach did not take interaction of aircraft in a group into account significantly. In discussion they underlined that interaction and interdependence are indeed main points of concern when systems of systems are evaluated for strengths and weaknesses, supporting the proposition of a mission result-based assessment in this work. Generally, the SMEs were very open towards the method presented here and expressed appreciation for the structured nature of the process.

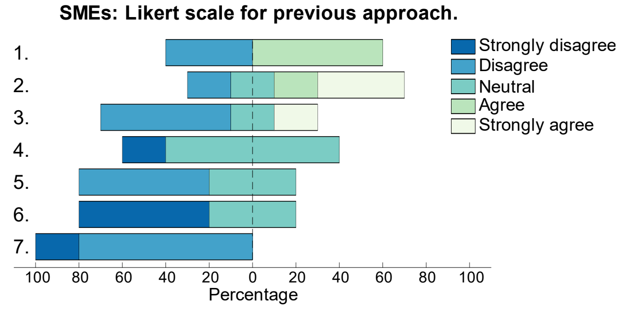

Figure 2: SMEs about the necessity of mission-based assessment for systems of systems and their previous approach on five-point Likert scale. For the statements refer to table 2.

6.1. Criteria Decomposition by SMEs

Figure 3 shows the result (upper levels) of the hierarchical decomposition of the demo scenario. Benefits or positive criterions (σ=+1) are the effect chain – which is directly connected to the primary goal of neutralizing the enemy convoy – and experience gained ("XP"). Effort criterions or negatives (σ=-1) are monetary cost and several measures of time. Side effects are ambivalent, as collateral damage is highly undesirable, but positive windfalls (e.g. destruction of a SAM site in self-defense), would represent benefits. For gaining the efficiency rating, the weighted sum over all benefits will be divided by the weighted sum of all effort criterions, as mentioned in chapter 4.

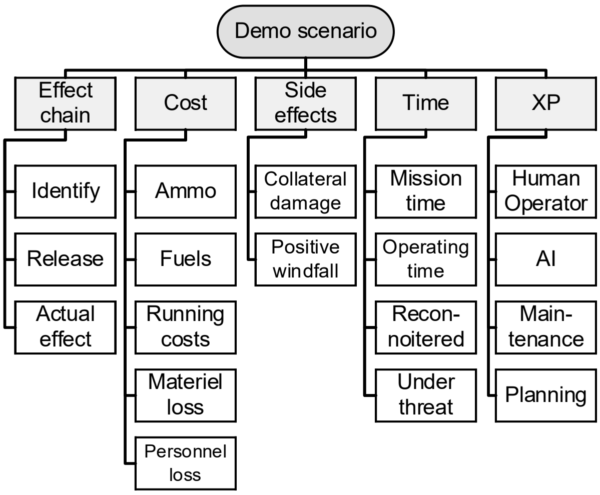

Figure 3: First two levels of decomposition of the example mission in a tree structure by the SMEs. "XP" means experience, "AI" artificial intelligence. "Fuels" also include all other operating fluids.

The brainwriting stage took about 15 minutes quite homogeneously among the SMEs, but no time target had been set by the moderator. The SMEs took notes in varying ways and some used structures such as mind maps. The results showed great differences, as was discovered in the subsequent group discussion.

From the beginning of the discussion (total duration of about two hours) it was highlighted that it is extremely important, but also quite time-consuming, to find accurate and universally agreed-upon definitions. This was reflected in multiple iterations of refining the definition of an item such as cost, and/or decomposing it further.

In the group discussion it was revealed that a mission planning perspective was used by some SMEs, while others wanted to be "thinking from the [desired] end state". The moderator rarely had to remind participants that they should think in more general and abstract terms for higher hierarchical levels and only become more specific in the criteria at the lower levels. Some criteria turned out to be heavily dependent on political and military constraints (which can be expressed by the verbs "want", "can", "be allowed to") – one must pay attention to this when defining the considered scenario.

Whereas the resulting upper two levels of the decomposition tree are rather general headlines, the actual measurable criteria were to be found mostly on the third level of decomposition.

Notable exception are the *times*. Among these operating time (engine running) is longer than the actual total mission time (from takeoff to landing). Then there is the time when an aircraft (part of the systems of systems) is detected by the opponent, i.e. is not covert. This can even be the case outside the area of operation. Probably the shortest time would be the time under threat, which is assumed to be only the case in the area of operation due to the presence of SAMs/MANPADS. Additional weighting of the last two types of time can be introduced based on severity and type of threat. A candidate metric for this is the reflected radar or IR signature integrated over the total simulation time.

Concerning the effect chain, it is obvious that there are severe correlations and interdependencies, namely without proper target acquisition and identification there will be no effector release, and without effector release there will be no impact. However, it was determined that a correct target acquisition and identification yields a (minor) benefit, even if no successful release or effect occurs. The actual effect on convoy in the demo scenario is rather easy to define: its functionality is inhibited if a certain percentage of vehicles cannot move any more. This idea of looking at how much of the functionality of the target is impeded is a bit harder to use but still useful for e.g. a terrorist camp. Importantly, according to SME discussion said functionality must be well defined, and sometimes can be hard to quantify, so an arbitrary scale or discrete values (0 and 1) could be applied.

Cost generally was expressed as monetary, but personnel loss cannot ethically be quantified in terms of monetary cost and its value must therefore be assumed to be either zero (no personnel loss) or negative infinity ($\sigma j_{personell\text{-}loss}=-\infty$). Materiel loss (e.g. of an UAV) on the other hand can be quantified by the cost of replacement.

After the decomposition itself, the optimization targets, mainly minimization and maximization, were quite easily found. A little discussion was required regarding the effect chain: While for ground target acquisition/identification and effector release the desired value has to be fixated, for the actual effect, the SMEs agreed that concerning the disabling of a convoy, below a threshold of e.g. 80% of the vehicles, there is a linear rise in benefit. Above this threshold there would be a less steep linear increase up to the maximum of 100%. A helpful question, mainly for finding thresholds, is “is there a higher utility if more/less of this happens?”

Normalization required some more discussion. It is necessary for all criteria evaluations to have a range from 0 to 1, in order not to inadvertently weight themselves. However, for monetary cost it might be better to replace the AHP-based weighting by using actual cost values in terms of Dollars or Euros etc. On the other hand, some SMEs mentioned that different types of costs indeed sometimes are seen and therefore weighted differently.

6.2. Weighting by SMEs

After they agreed on the decomposition structure, each individual SME completed the comparison matrices for all decomposition nodes fully independently before the second group discussion stage. After said hierarchical structure was loaded into the software (see 5.2), the moderator only offered methodical guidance, i.e. gave an introduction on how to use the GUI. The GUI turned out to indeed be easy to use as intended, each SME taking between five and ten minutes for completing all comparison matrices.

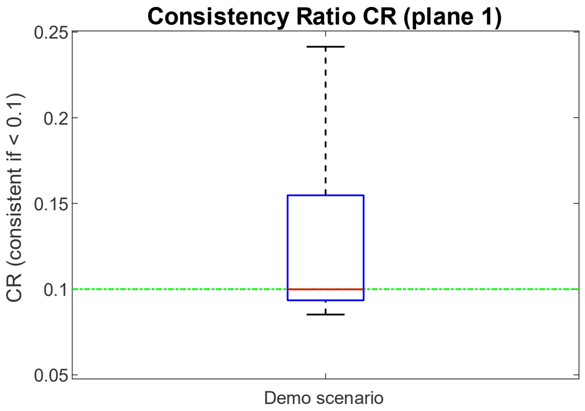

Figure 4: Consistency ratio CR for top node after second round of FAHP. The dotted line marks the acceptable limit of 0.1.

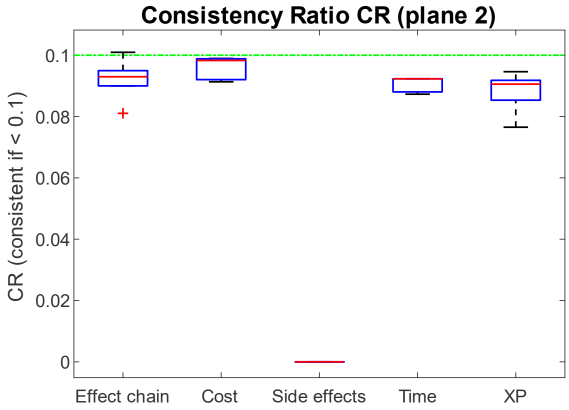

Figure 5: Consistency ratios CR for nodes on first level of decomposition after second round of FAHP as boxplots for all SMEs. The dotted line marks the acceptable limit of 0.1.

It was observed that in the first round most SMEs relied on a rather intuitive approach for choosing the linguistic variables (cf. table 1). Difficulties arose when weighing criteria with correlations or interdependencies against each other, which is the case regarding the effect chain. Over the whole FAHP stage, only one SME was shortly contemplating the use of the Simos option, but in the end did not make use of it, although it was explained and offered to all SMEs. Four out of five SMEs did not achieve a CR under 0.1 at the first round of FAHP.

When being presented with the results of the first round of FAHP, discussion ensued among the experts, especially about the sub-criteria of “side effects”. Two SMEs had rated “positive windfalls” as highly more important than collateral damage. It soon turned out that they only had different understandings of the meaning of the linguistic variables when using the program: For all SMEs really no collateral damage was viewed as acceptable when human lives are concerned, however to some degree unwittingly damaging items could be accepted. These two SMEs had obviously flipped the sign σ unconsciously; they explained that the *avoidance* of collateral damage is highly more important than any positive windfalls could be. So, whereas all SMEs were indeed in agreement on this issue, they reacted

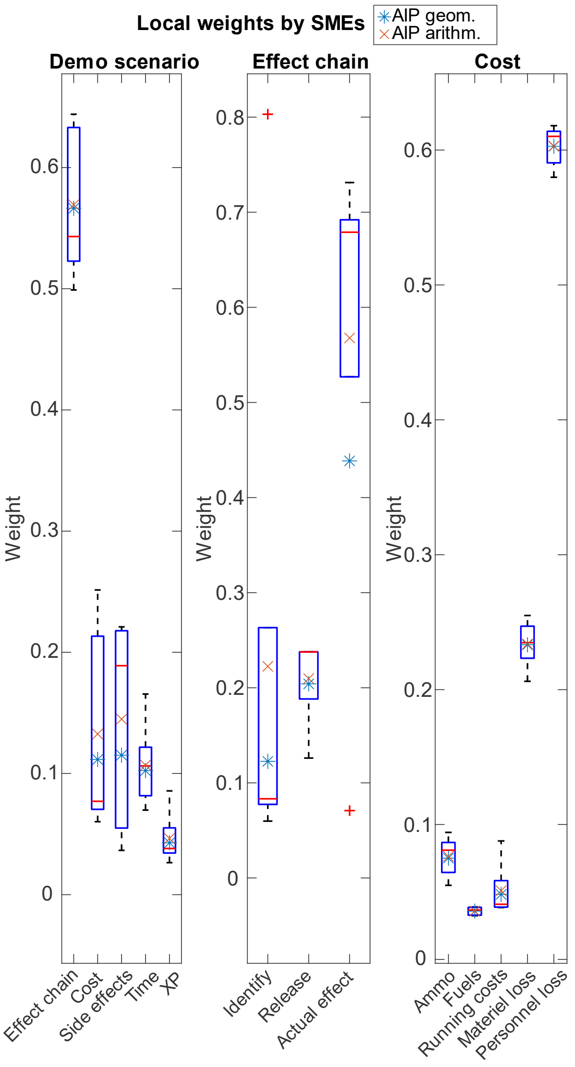

Figure 6: Local weights by all SMEs after second round of FHAP as boxplots. Aggregation via AIP with arithmetic mean (star symbol) and geometric mean (x symbol).

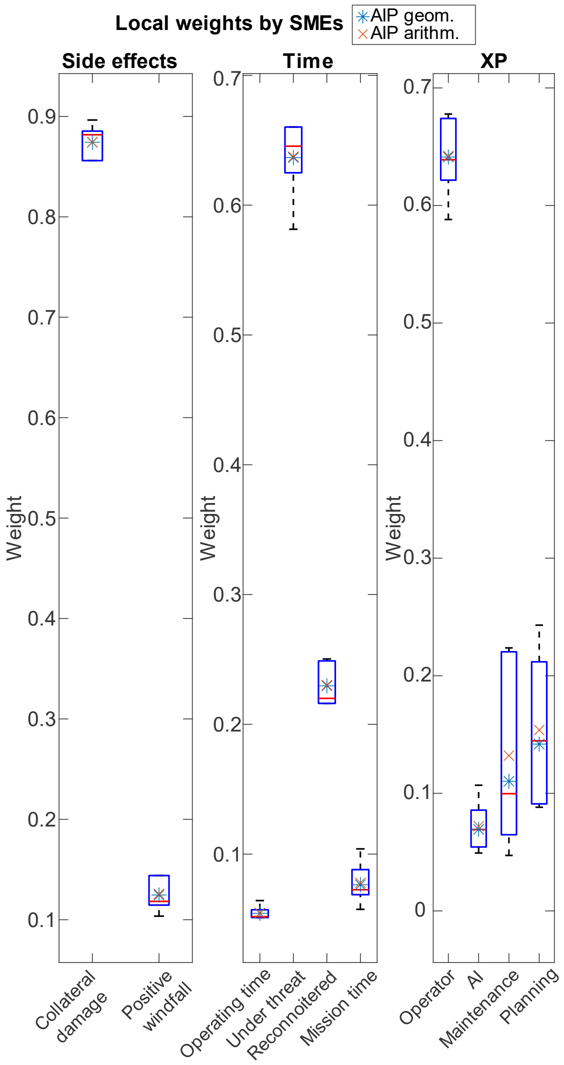

Figure 7: Local weights by all SMEs after second round of FHAP as boxplots. Aggregation via AIP with arithmetic mean (star symbol) and geometric mean (x symbol).

differently in terms of weighting before the language was refined. While there was overwhelming group agreement on some nodes, quite different importance ratings were found on other nodes. An attempt to cluster SMEs revealed that there was no single SME who was more consistent (lower CR) on all decomposition nodes, nor were there any constant groups of SMEs with roughly the same opinions on all nodes.

The relatively high CRs required one iteration loop with slight tweaks by each SME on their chosen linguistic variables for single items. Only minor modifications of comparison matrices were made to gain CR lower than 0.1, but these resulted in no changes in the ranking order of criteria at their respective node, just small differences in the distances.

Figures 4 and 5 show the CRs for the mission tree nodes from figure 3 after the second FAHP round. For the top node “Demo scenario” with its five sub-criteria the median CR is slightly below 0.1. One outlier is well above even 0.2 and therefore this SME should do another FAHP iteration at this decomposition node. For the comparison matrices of the elements of the five sub-criteria (“Effect chain” etc.) however consistency is acceptable. Note that the node “side effects” automatically is consistent ($CR=0$), because there are only two items in its comparison matrix.

For group aggregation the method AIP was chosen, because it is easy to implement and explain to the SMEs to ensure transparency. Also it fulfills the Pareto optimality axiom [23] and is recommended over AIJ when the SMEs act as individuals [24] (which is an assumption supported by the fact they could not be clustered into subgroups). LFA was not as transparent as AIJ or AIP to the SMEs and was therefore dismissed, notwithstanding the also present issue with LFA if there are only two elements in a node (as was the case with “side effects”), because then no finite consistency ratio is available.

Figures 6 and 7 depict the local weights of all criteria mentioned in figure 3 as boxplots (including the median) representing the SMEs choices. Small boxes mean a low spread in the opinion of the SMEs regarding the importance of the criterion, taller boxes mean differing weightings by SMEs. Especially concerning cost, side effects, and time, there obviously is a quite good consensus among the experts. For the total demo mission, the effect chain and “XP” the opinions were more diverse.

The node “effect chain” shows one SME marked as outlier, who basically interchanged the importance of “identification” and “actual effect” compared to the other participants. The reason given in discussion was that without proper target acquisition and identification there should and would be no effector release and therefore no effect. The other interviewees maintained that for a mission assessment they regarded the additional benefit of a successful effect on the target as significantly higher.

Also shown in figures 6 and 7 are the aggregation results using AIP with the arithmetic mean (star symbol) and the geometric mean (x symbol) methods. Note that the aggregated values should be normalized again, so that the sum of all local weights at one node gives one. Use of the geometric mean with AIP is suggested, because it fulfills the reciprocal property of AHP [23, 24]. When the spread among the SMEs is small, both lie rather closely together.

After aggregation and calculation of the total weights the explicit objective function for the overall mission effectiveness for the specific scenario can be constructed as seen in equation 1. An individual assessment of the mission can be conducted for each SME when only their respective weighting results are used in equation 1 instead of the aggregated weights. The mission efficiency value is gained by dividing the effectiveness J by the sum of all weighted cost criteria.

6.3. Method evaluation by SMEs

While the first part of the questionnaire (table 2, figure 2) highlights the need and requirements for a new method, the second part was for evaluation of the presented method itself. The seven statements in table 3 were responded to by the SMEs again on a five-point Likert scale indicating graduated agreement or disagreement. The results are shown in figure 8, while the right column in table 3 gives an indication whether the SMEs’ answers were generally favorable to the presented method.

Almost all SMEs indicated they would like to incorporate the demonstrated method in their future analyses or assessments. The SMEs unanimously agreed that the

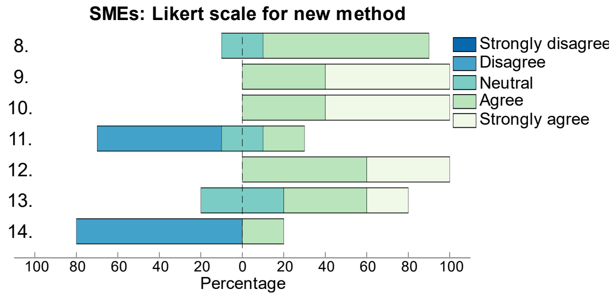

Figure 8: SMEs about the presented new method, on five-point Likert scale. For the statements refer to table 3.

demonstrated method makes sense and that it is well structured. That hints at the process being significantly more structured than the previous approaches. Most SMEs at least slightly prefer the new method compared to their previous approach of judging aircraft missions. All surveyed SMEs agree or strongly agree that the demonstrated method includes relevant criteria more completely than their previous approach. This can be attributed to the higher attained level of detail due to the methodical, level-by-level decomposition and the feedback loops during the process. Regarding the effort this process takes, the SMEs rather agreed that it is justified in relation to the benefit. 80% of the SMEs also did not see the effort as too high in relation to the obtained increase in detail.

In summary, the presented method was evaluated generally very favorably by the SMEs.

#	Statement	Answers favorable +/o/-
8	I would like to use the demonstrated method as an element of future analyses/assessments.	+
9	I think the demonstrated method makes sense.	+
10	I think the demonstrated method is well structured.	+
11	I prefer my previous approach.	+
12	The demonstrated method includes relevant criteria more completely than the previous approach.	+
13	The effort is justified in relation to the benefit.	+
14	The effort is too high for the obtained increase in detail.	+

Table 3: Evaluation of the new method: Statements presented to SMEs and favorability of their answers towards the demonstrated method.

7. CONCLUSION AND FUTURE WORK

The results of the SME questionnaire indicate the presented AHP-based approach should indeed be further pursued, because an appropriate method for assessment of both missions and systems-of-systems (composites of aircraft) is needed. It delivers more detail and structure than previously used approaches by the systematic hierarchical decomposition of a concrete mission, the iterative use of FAHP for weighting, and the overall incorporation of various SMEs' knowledge. In the group discussions special attention to nuances in language and using universally agreed definitions is required.

This work is a first step in the development of a new assessment methodology for aerial systems of systems. In further research the last two stages of the concept (figure 1) must be implemented and checked for viability:

Actual simulation of the chosen demo scenario and subsequent generation of aggregated assessments for different systems will be undertaken in a multi-agent simulation, that can represent various types of interactions of the individual aircraft in a composite force. By comparing different combined forces, e.g. with different numbers of members in the force package or varying mission equipment such as sensors, it can be shown that the method indeed highlights differences between distinct aircraft composites.

For validation, the original five SMEs will have to compare the final aggregated numerical measure of effectiveness J, a "dashboard"-style mission overview, and more detailed mission run information. With their experience, intuition and previous approaches they need to determine whether the assessment by the proposed method is believable and accurate to their individual and combined input to the metric derivation process.

Also for validation, local and global sensitivity analysis of weights and measurement results is very important, because significant differences in perceived mission execution must yield noticeable changes in the overall measure of effectiveness J.

Further research then is necessary in the area of *Design of Experiments* (DoE) for sampling stratification. When multiple missions have to be simulated in many different ways, e.g. varying plans, for generating an overall assessment of a system of systems, unacceptably long computation times must be avoided. In the same vein, it must be determined what degree of detail is required in simulation to capture all relevant interaction effects but is still viable in terms of computation time.

REFERENCES

[1] D. N. Mavris, D. S. Soban, and M. C. Largent, "An Application of a Technology Impact Forecasting (TIF) Method to an Uninhabited Combat Aerial Vehicle," in *SAE Technical Paper Series*, 1999.

[2] D. S. Soban, "A Methodology for the Probabilistic Assessment of System Effectiveness as Applied to Aircraft Survivability and Susceptibility," Ph.D. dissertation, Georgia Institute of Technology School of Aerospace Engineering, Atlanta, Georgia, 2001.

[3] P. Stodola, J. Drozd, J. Mazal, J. Hodický, and D. Procházka, "Cooperative Unmanned Aerial System Reconnaissance in a Complex Urban Environment and Uneven Terrain," *Sensors (Basel, Switzerland)*, vol. 19, no. 17, 2019, doi: 10.3390/s19173754.

[4] N. K. Jaiswal, *Military operations research: Quantitative decision making*. Boston, Mass.: Kluwer Academic Publishers, 1997.

[5] A. R. Habayeb, *Systems Effectiveness*. Burlington: Elsevier Science, 1987. [Online]. Available: http://gbv.eblib.com/patron/FullRecord.aspx?p=1829225

[6] R. J. Hillestad and L. Moore, "The Theater-Level Campaign Model: A Research Prototype for a New Generation of Combat Analysis Model," RAND Corp., Santa Monica CA RAND-MR-388-AF/A, 1996. Accessed: May 6 2020. [Online]. Available: https://

apps.dtic.mil/docs/citations/ADA319651

[7] D. Kallfass and T. Schlaak, “NATO MSG-088 Case Study Results to Demonstrate the Benefit of Using Data Farming for Military Decision Support,” *Proceedings of the 2012 Winter Simulation Conference*, pp. 2481–2492, 2012. [Online]. Available: https://informs-sim.org/wsc12papers/includes/files/con502.pdf

[8] G. Horne ***et al.***, “Developing Actionable Data Farming Decision Support for NATO: Final Report of MSG-124,” NATO AC/323(MSG-124)TP/825, 2018. Accessed: Mar. 30 2020. [Online]. Available: https://www.foi.se/download/18.7fd35d7f166c56ebe0be163/1542369114003/Developing-Actionable-Data-Farming-Decision-Support-for-NATOTR_STO-TR-MSG-124.pdf

[9] N. D. Ernest, “Genetic Fuzzy Trees for Intelligent Control of Unmanned Combat Aerial Vehicles,” College of Engineering and Applied Science, University of Cincinnati, 2015.

[10] S. Morawietz, M. Strohal, and P. Stütz, “A Mission-Based Approach for the Holistic Evaluation of Aerial Platforms: Implementation and Proof of Concept,” in *18th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference 2017: Denver, Colorado, USA, 5-9 June 2017 : held at the AIAA Aviation Forum 2017*, Denver, Colorado, 2017.

[11] S. Morawietz, M. Strohal, and P. Stütz, “A Decision Support System for the Mission-Based Evaluation of Aerial Platforms: Advancements and Final Validation Results,” in *18th AIAA Aviation Technology, Integration, and Operations Conference 2018: Atlanta, Georgia, USA, 25-29 June 2018 : held at the AIAA Aviation Forum 2018*, Atlanta, Georgia, 2018.

[12] C. Ruf, M. Zwick, S. Morawietz, and P. Stütz, “Enhancing Automated Aerial Reconnaissance Onboard UAVs Using Sensor Data Processing-Characteristics and Pareto Front Optimization,” in *AIAA Scitech 2019 Forum*, San Diego, California, 2019.

[13] U. Buscher, A. Wels, and R. Franke, “Kritische Analyse der Eignung des Fuzzy-AHP zur Lieferantenauswahl,” in *Supply Management Research: Aktuelle Forschungsergebnisse 2010; [Tagungsband des wissenschaftlichen Symposiums Supply Management; Advanced studies in supply management, Bd. 3]*, R. Bogaschewsky, M. Eßig, R. Lasch, and W. Stölzle, Eds., 1st ed., Wiesbaden: Gabler, 2010, pp. 27–60.

[14] A. B. vanGundy, “BRAINWRITING FOR NEW PRODUCT IDEAS: AN ALTERNATIVE TO BRAINSTORMING,” *Journal of Consumer Marketing*, vol. 1, no. 2, pp. 67–74, 1983. [Online]. Available: https://ezproxy.bibl.unibw-muenchen.de/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edb&AN=8941179&lang=de&site=eds-live

[15] R. W. Saaty, “The analytic hierarchy process—what it is and how it is used,” *Mathematical Modelling*, vol. 9, no. 3, pp. 161–176, 1987, doi: 10.1016/0270-0255(87)90473-8.

[16] E. Mu and M. Pereyra-Rojas, *Practical Decision Making: An Introduction to the Analytic Hierarchy Process (AHP) Using Super Decisions v2*. Cham: Springer International Publishing, 2017.

[17] I. N. Bronštejn, K. A. Semendjaev, G. Musiol, and H. Mühlig, *Taschenbuch der Mathematik,* 9th ed. Haan-Gruiten: Verl. Europa-Lehrmittel, 2013.

[18] T. L. Saaty and G. Hu, “Ranking by Eigenvector versus other methods in the Analytic Hierarchy Process,” *Applied Mathematics Letters*, vol. 11, no. 4, pp. 121–125, 1998, doi: 10.1016/S0893-9659(98)00068-8.

[19] W. C. Wedley, “Consistency prediction for incomplete AHP matrices,” *Mathematical and Computer Modelling*, vol. 17, 4-5, pp. 151–161, 1993, doi: 10.1016/0895-7177(93)90183-Y.

[20] R. Csutora and J. J. Buckley, “Fuzzy hierarchical analysis: the Lambda-Max method,” *Fuzzy Sets and Systems*, vol. 120, no. 2, pp. 181–195, 2001, doi: 10.1016/S0165-0114(99)00155-4.

[21] J. Krejčí, “Fuzzy eigenvector method for obtaining normalized fuzzy weights from fuzzy pairwise comparison matrices,” *Fuzzy Sets and Systems*, vol. 315, pp. 26–43, 2017, doi: 10.1016/j.fss.2016.03.006.

[22] E. Siskos and N. Tsotsolas, “Elicitation of criteria importance weights through the Simos method: A robustness concern,” *European Journal of Operational Research*, vol. 246, no. 2, pp. 543–553, 2015, doi: 10.1016/j.ejor.2015.04.037.

[23] W. Ossadnik, S. Schinke, and R. H. Kaspar, “Group Aggregation Techniques for Analytic Hierarchy Process and Analytic Network Process: A Comparative Analysis,” *Group Decis Negot*, vol. 25, no. 2, pp. 421–457, 2016, doi: 10.1007/s10726-015-9448-4.

[24] E. Forman and K. Peniwati, “Aggregating individual judgments and priorities with the analytic hierarchy process,” *European Journal of Operational Research*, vol. 108, no. 1, pp. 165–169, 1998, doi: 10.1016/S0377-2217(97)00244-0.

[25] Y.-G. Cho and K.-T. Cho, “A loss function approach to group preference aggregation in the AHP,” *Computers & Operations Research*, vol. 35, no. 3, pp. 884–892, 2008, doi: 10.1016/j.cor.2006.04.008.

[26] R. Likert, “A technique for the measurement of attitudes,” *Archives of Psychology*, 22 140, p. 55, 1932.

APPENDIX

n	1	2	3	4	5	6	7	8	9	10
RI	0	0	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49

Table 4: Random consistency index RI for a scale 1/9,1/8,…,8,9 according to [15].

Corresponding author

julian.seethaler@unibw.de