# Validating Air Combat Behaviour Models for Adaptive Training of Teams

Armon Toubman(✉)

Netherlands Aerospace Centre NLR, Amsterdam, The Netherlands
Armon.Toubman@nlr.nl

**Abstract.** On many occasions, the use of machine learning to adaptively generate new content for training simulations has been demonstrated. However, the *validation* of the new content (i.e., proof that the new content is *fit for use* in training simulations), has received relatively little attention. In this study, we design a validation procedure for one particular type of content, namely the behaviour models for the virtual opponents in air combat training simulations. As a case study, we generate a new set of behaviour models and apply the validation procedure to them. Our results are positive, but leave room for interpretation. We discuss why this is the case and suggest avenues for future work.

**Keywords:** Adaptive training · Air combat · Model validation

## 1 Introduction

In air combat training simulations, the role of opponent is often played by virtual entities known as computer generated forces (CGFs). Various research efforts have demonstrated the ability of machine learning (cf. Karli et al. 2017; Teng et al. 2013; Toubman et al. 2016) and other adaptive techniques (cf. Floyd et al. 2017; Karneeb et al. 2018) to generate air combat behaviour models for CGFs. The strength of such techniques is that the computer can automatically adapt the behaviour of the CGFs, and thus the training, to the trainee fighter pilots. However, the creative capabilities of these techniques may result in undesirable (e.g., non-humanlike) behaviour that is not useful for training (Petty 2003). The main idea behind this paper is that newly generated behaviour models should be validated to prove their usefulness in training simulations. In the remainder of this paper, we investigate what this validation entails (see Sect. 2). The two contributions of this paper are the following:

1. We present a validation procedure for machine-learned air combat behaviour models (see Sect. 3). A key component of the procedure is a newly developed questionnaire for the assessment of the behaviour produced by air combat behaviour models. We call this questionnaire the Assessment Tool for Air Combat CGFs (ATACC);


R. A. Sottilare and J. Schwarz (Eds.): HCII 2019, LNCS 11597, pp. 557–571, 2019.
https://doi.org/10.1007/978-3-030-22341-0_44

2. As a case study, we generate novel air combat behaviour models by means of machine learning (see Sect. 4) and apply the validation procedure to the models (see Sect. 5). The results show that the generated behaviour models are valid to some extent, but also that both the behaviour models and the validation procedure require additional effort (see Sect. 6). To the best of our knowledge, this is the first time that the validation of machine-generated air combat behaviour models has been treated as a research subject in its own right (see Sect. 7).

## 2 The Difficulty of Validating Behaviour Models

Since the advent of the use of simulation in military training there has been a rising interest in the *validation* of simulation models (cf. Kim et al. 2015; Sargent 2011). Many definitions of validation have been stated throughout the literature (cf. Birta and Arbez 2013; Bruzzone and Massei 2017; Petty 2010). When military simulations are discussed in particular, we find references to the definition of validation that is used by the US Department of Defense (2009). We use this definition from now onwards. For convenience, we restate the definition.

**Definition 1 (Validation).** *Validation is "[t]he process of determining the degree to which a model or simulation and its associated data are an accurate representation of the real world from the perspective of the intended uses of the model" (ibid.).*

The definition names four important concepts: (1) a process, (2) a degree of accuracy, (3) a model (or simulation), and (4) the intended use of the model. We can readily fill in concepts (3) and (4). Regarding concept (3), the models that we wish to validate are newly generated behaviour models. Furthermore, regarding concept (4), the intended use of these models is to produce behaviour for opponent CGFs in air combat training simulations. However, this leaves open two questions for us to investigate: (1) what the process entails, and (2) how we should determine the accuracy of the models. *The difficulty of validating behaviour models is answering these two questions for every specific case.*

First, we investigate the question of *what the process entails*. There is no *one-size-fits-all* solution for validation processes, since different models have (1) different intended uses, and (2) different *associated works* available for use in the validation. Here, we use the notion "associated work" to refer to a range of results of performed work, e.g., (1) baseline models, (2) expected output data, (3) conceptual diagrams of the modelled phenomenon, or (4) expert knowledge. This being so, we still find that the various validation methods to be applied are well described in the literature (cf. Balci 1994; Petty 2010; Sargent 2011). In general, the four categories of validation methods are: (1) informal methods such as face validation, (2) static methods such as evaluating the model structure, (3) dynamic methods that involve executing the model and analysing the output data, and (4) formal methods based on mathematical proofs. An important factor in the choice of validation method(s) to use is the availability of associated

works (cf. Petty 2010; Sargent 2011). For example, dynamic methods can only be applied if (1) it is possible to execute the model with input that is relevant with regard to the intended use of the model, (2) data can be collected on the execution of the model, and (3) it is known how the collected data should be interpreted (e.g., compared to another available set of data). In other words, the choice of validation methods is always limited by practical considerations.

The second question we would like to investigate reads: *how should we determine the accuracy of the models?* For instance, for a physics-based model, the accuracy of the model can be defined in terms of the number of faults that is allowed when the data that the model produces is compared to data that is measured in the real world. However, for behaviour models the question is particularly difficult to answer, since the notion of fault is difficult to grasp (Hahn 2017). Goerger et al. (2005) identify five causes to the difficulty of validating behaviour models in general. Four[1] of these causes relate to the problem of defining the accuracy of a behaviour model. These four causes are: (1) the cognitive processes that are modelled may be nonlinear, which makes the processes as well as their models hard to reason about, (2) it is impossible to investigate all possible interactions that may arise in simulations because of the large number of interdependent variables in the models, (3) the metrics for measuring accuracy are inadequate, (4) there is no "robust" set of input data for the models.

An important consequence of the difficulty of validating behaviour models is that the outcome of a validation should not be interpreted as either "the model is valid" or "the model is not valid", as it is practically impossible to "completely validate" a model (Birta and Arbez 2013). Therefore, Birta and Arbez, (ibid.) note that "degrees of success must be recognized and accepted." For them, it is important that the chosen validation methods are able to adequately reflect on the extent of the validity of the models.

## 3 Our Proposed Validation Procedure

In this section, we present our validation procedure for air combat behaviour models. Specifically, the validation procedure is aimed at automatically generated (e.g., machine-learned) behaviour models. The main idea behind the validation procedure is a comparison of (a) the behaviour displayed by CGFs that use the generated behaviour models, to (b) the behaviour displayed by CGFs that use behaviour models that have been written by professional model builders and/or subject matter experts (henceforth the *professionals*). Essentially, we use the latter, *established* type of behaviour models to provide a standard of behaviour to which the former, *newly generated* type of behaviour models should adhere. In other words, we do not aim for the generated models to *surpass* the established models in any way. Rather, we aim to show their equivalence, so that the new models can be used to supplement the established models, and thereby widen the variety of the training simulations that are offered.

[1] The fifth cause is the lack of a standard validation process, which we discussed earlier.

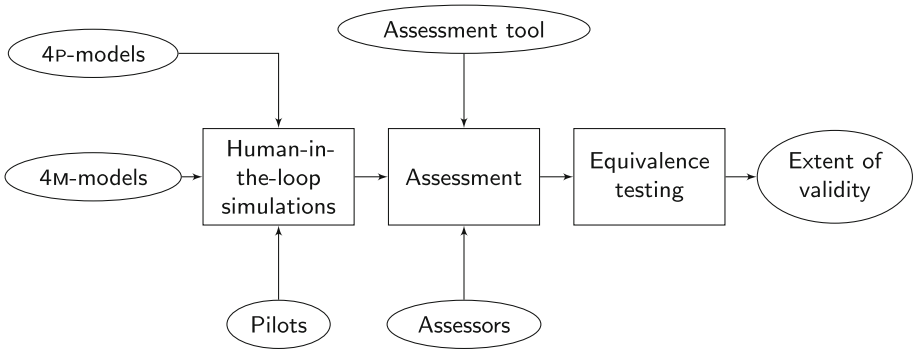


**Fig. 1.** The validation procedure. In human-in-the-loop simulations, human fighter pilots engage CGFs that are either controlled by the 4M-models (subject of the validation) or the 4P-models (baseline for comparison). Expert assessors assess the behaviour displayed by the CGFs by means of a newly developed assessment tool. Equivalence testing on the assessment results in a measurable extent of validity of the 4M-models.

In order to produce observable (and thus comparable) behaviour, all of the models have to be fed with the behaviour of their opponents, i.e., CGFs controlled by human fighter pilots, in a realistic air combat setting (see Sect. 3.1). Next, the displayed behaviour has to be assessed to create data on the basis of which a comparison can be made (see Sect. 3.2). For the actual comparison, we rely on a statistical method known as *equivalence testing* (see Sect. 3.3). Based on the outcome of the equivalence testing, we can state the extent of the validity of the generated behaviour models. Figure 1 provides an overview of the entire validation procedure.

### 3.1 Human-in-the-Loop Simulations

The validation procedure begins with human-in-the-loop simulations in a high-fidelity beyond-visual-range air combat simulator. We consider a simulator that accommodates four human participants acting as fighter pilots. In the simulations, the participants engage a so-called *four-ship* (viz. a team of four) of hostile CGFs.

The behaviour of the four-ship of CGFs is driven by four behaviour models, one for each CGF. In our experience, the behaviour models for the CGFs in a four-ship are treated as a single model. Especially when the models are designed by professionals, they are carefully tuned to each other to provide the illusion of a cohesive team at work. We henceforth consider the four models that together control the behaviour of a four-ship to be an indivisible unit. For convenience, we introduce the term *4-model* to refer to a group of four behaviour models.

Using the term 4-model, we are now able to make the distinction between (1) 4-models that have been written by the professionals, and (2) 4-models that have been generated by means of machine learning. We introduce the terms 4P-model

(where the P stands for *professional*) and 4M-model (where the M stands for *machine learning*) to refer to these two kinds of 4-model, respectively.

The 4M-models are the subjects of the validation procedure. However, by themselves they are not sufficient input for the validation process. As Petty (2010) stated succinctly, validation "[is a] process[] that compare[s] things." Therefore, we require either (a) a baseline model, (b) a set of expected output data, or (c) implicit expert knowledge as a reference to compare against the 4M-models.

For complex air combat behaviour models, it is almost infeasible to compile a set of expected output data, since the output depends on a wide range of possible interactions with other entities. However, what we *do* have available are behaviour models that have been written previously by professionals (i.e., 4P-models). These 4P-models constitute a *sample* of all behaviour models that have been written by the professionals, comparable to how the 4M-models that are validated are a sample of the behaviour models that can possibly be generated by machine learning. Furthermore, we argue that since the 4P-models have been developed by means of the behaviour modelling process, the 4P-models have themselves been validated to some extent. We therefore add 4P-models as the second input to the validation process.

We record the human-in-the-loop simulations, resulting in a set of *behaviour traces*. The behaviour traces contain three-dimensional recordings of the simulated airspace, including the movements of all entities (i.e., CGFs and missiles) flying in the airspace. The behaviour traces serve as input for the assessment (see next section).

### 3.2 Assessment

The goal of the assessment is to summarise the CGFs' behaviour that is encoded in the behaviour traces into values that are (1) meaningful and (2) comparable between the 4M-models and the 4P-models. The assessment is performed by means of a structured form of face validation, which is one of the informal validation methods (see Sect. 2).

However, there is little to no information available on measures for CGF behaviour that are relevant to training simulations. Therefore, we make use of the implicit knowledge of expert evaluators. We leverage this knowledge in two manners. First, we elicit knowledge on measures for behaviour of air combat CGFs, and then structure this knowledge into a novel assessment tool which we call the *Assessment Tool for Air Combat* CGFs (ATACC) (see below). This tool enables a structured assessment of CGF behaviour. Second, expert evaluators review the behaviour traces that we have collected, and then assess the behaviour that the CGFs display. The result of the assessment is a series of ratings on Likert scales. The ratings serve as input for the equivalence tests (see next section). Below, we describe the development and contents of the ATACC.

**The Assessment Tool for Air Combat CGFs.** Together with instructor fighter pilots, we identified three performance dimensions that should be taken into consideration in the assessment of the behaviour of air combat CGFs. These performance dimensions are (1) the *challenge* provided by the CGFs, (2) the *situational awareness* that the CGFs display, and (3) the *realism* of the behaviour of the CGFs. We briefly describe the three performance dimensions below.

**Performance dimension 1: Challenge.** The tool should measure whether (1) the CGFs behave in such a way that the human participants in the simulations need to think about and adjust their actions, and (2) whether the CGFs provide some form of *training value* to the simulations.

**Performance dimension 2: Situational awareness.** The tool should measure whether (1) the CGFs appear to sense and react to changes in their environment, and (2) whether multiple CGFs belonging to the same team appear to acknowledge each other's presence.

**Performance dimension 3: Realism.** The assessment tool should measure (1) whether the CGFs behave as can be expected from their real-world counterparts, and (2) whether the CGFs use the capabilities of their platform (including e.g., sensors and weapons) in a realistic manner.

Next, we attempted to formulate examples of behaviour that relate to each of the performance dimensions. This was done in an iterative manner, such that examples that were proposed could be critically analysed by each of the instructor fighter pilots. We formulated eight examples of behaviour in total (listed below). Examples 1 through 4 relate to *Challenge*; 5 and 6 to *Situational awareness*; and 7 and 8 to *Realism* In each of the examples, *red air* refers to the CGFs, whereas *blue air* refers to the human participants in the human-in-the-loop simulations.

**Example 1.** Red air forced blue air to change their tactical plan.
**Example 2.** Red air forced blue air to change their shot doctrine[2].
**Example 3.** Red air was within factor range[3].
**Example 4.** Blue air was able to fire without threat from red air.[4]
**Example 5.** Red air acted on blue air's geometry.
**Example 6.** Red air acted on blue air's weapon engagement zone[5].
**Example 7.** Red air flew with kinematic realism.
**Example 8.** Red air's behaviour was intelligent.

In the ATACC, each of the eight examples of behaviour is presented as a separate rating item, so that the presence of each behaviour is rated on a five-point Likert scale. For all of the eight rating items, the scale is labelled as ranging

[2] Jargon: pre-briefed instructions for the use of air-to-air weapons.
[3] Jargon: the range within which opponents have to be taken into account in the selection of tactical actions.
[4] We formulated this behaviour from the viewpoint of blue air, since we were unable to satisfactorily state the behaviour from the viewpoint of red air.
[5] Jargon: the airspace in front of a fighter jet in which a fired missile can be effective.

from *Never* to *Always*. To conclude the ATACC, we added a general ninth rating item stating "Red air's behaviour tested blue air's tactical air combat skills." This item served to provide us with a general indication of the usefulness of the behaviour of the CGFs in relation to the human-in-the-loop simulations that were performed. The ninth item is also rated on five-point Likert scale, ranging from *Strongly disagree* to *Strongly agree*.

### 3.3 Equivalence Testing

At this point in the validation process, we have two sets of data: (1) the assessment of the 4P-models, and (2) the assessment of the 4M-models. We wish to compare these two sets of data in a meaningful way. Since we used the 4P-models as the baseline, we assume that the assessment of the 4P-models contains information about the desirable properties of air combat CGF behaviour. Based on this assumption, we define the measure of validity of the 4M-models as the extent that the assessment of the 4M-models and the assessment of the 4P-models can be measured to be equivalent.

Obviously, a simple comparison (viz., determining if the difference between the assessments equals zero) of the assessments is too strict. The results of our assessments include noise from multiple sources (e.g., the behaviour of the pilots in the human-in-the-loop simulations, and possible bias of the assessors). Furthermore, standard statistical significance tests do not suffice, since these tests check for differences rather than for equivalence. We found a solution in a form of comparison testing that is called *equivalence testing*.

The *two one-sided tests* (TOST) method tests for equivalence of the means of two populations (cf. Anderson-Cook and Borror 2016; Lakens 2017; Meyners 2012). Therefore, the method starts with the assumption that two populations are different, and then collects evidence to show that the populations are the same. Note that this is the opposite of traditional tests that compare two populations (e.g., Student's $t$-test), which (1) start with the assumption that two populations are similar or even the same, and then (2) collect evidence to show that the populations are different.

In TOST, the assumption that two populations are different (viz., the *null hypothesis* or $H_0$) is stated as follows.

$$H_0: \quad \mu_A - \mu_B \leq \delta_L \quad \text{or} \quad \mu_A - \mu_B \geq \delta_U \tag{1}$$

Here, the difference of the means of two populations A and B are compared. Two populations are considered different if the difference of their means lies outside of the indifference zone $[\delta_L, \delta_U]$. We assume that the indifference zone is symmetrical, i.e., $\delta = \delta_U = -\delta_L$. However, we are interested in examining the hypothesis viz. the means are not different, i.e., the difference between the means lies inside of the indifference zone. The reformulation of the hypothesis (viz. the alternative hypothesis or $H_{1)}$ is stated as follows.

$$H_1: \quad \delta_L < \mu_A - \mu_B < \delta_U \tag{2}$$

If the TOST finds evidence that the difference of the means lies within the indifference zone under the assumption that it does not, we reject $H_0$ and accept $H_1$, meaning that we conclude that the populations are the same (up to a very small difference). Finding this evidence is done by splitting $H_0$ into two hypotheses which can be tested using standard one-sided $t$-tests. The $p$-value of the TOST then becomes the maximum of the two $p$-values that are obtained from the two one-sided $t$-tests.

The outcome of the TOST greatly depends on the value chosen for $\delta$. Until recently, $\delta$ could not be calculated directly. It was either (1) prescribed by regulatory agencies (e.g., in the field of pharmacology) or (2) determined by subject matter experts based on reference studies or expectations about the data (e.g., in psychology) (cf. Anderson-Cook and Borror 2016; Lakens 2017). For our validation, it is difficult to determine a suitable $\delta$, since we have neither a regulatory agency, nor a reference study available. However, in 2016, an objective calculation of $\delta$ was introduced by Juzek (2016). The calculation of this delta $\delta$ (henceforth: Juzek's $\delta$) is as follows.

$$\delta = 4.58 \frac{s_p}{N_p} \tag{3}$$

Here, $s_p$ is the pooled standard deviation in the two samples under comparison, and $N_p$ is the pooled number of data points in the samples. Juzek found the coefficient (4.58) by simulating a large number of TOST applications. The coefficient was approximated in such a way that Juzek's $\delta$ gives the TOST the appropriate statistical power ($1 - \alpha = 95\%$, $1 - \beta = 80\%$).

Armed with the TOST method, we are now able to test the statistical equivalence of the assessments for the 4P-models and the 4M-models per rating item. The extent to which the rating items are equivalent can then be seen as the extent to which the 4M-models are valid.

## 4 Generating Air Combat Behaviour Models

We generated four novel 4M-models in preparation for the application of the validation procedure. These 4M-models served as the subject of the validation. The 4M-models were generated by means of the dynamic scripting machine learning algorithm (Spronck et al. 2006). The specific method for applying dynamic scripting to generate the 4M-models for air combat simulations is described by Toubman et al. (2016). We do not restate the full method here, as it is not the focus of this paper. In brief, the method consists of the following three steps:

1. We obtain four 4P-models that have been written by a professional and that have seen use in actual training simulations;
2. We decompose the 4P-models into their constituent "states"[6] and the transitions between these states;

[6] In our method, a state defines a "piece of behaviour", such as but not limited to "firing a missile" or "defensive moves".

3. The dynamic scripting algorithm repeatedly recombines the states and transitions into new behaviour models (4M-models) and tests these models in automated, agent-based simulations. The algorithm halts after a certain number of repetitions and returns the four best performing (viz. most-winning) 4M-models that it has found.

The use of this method thus results in (a) the four professionally written 4P-models obtained in the first step, and (b) the four machine-generated 4M-models obtained in the third step. Together, the eight models serve as input to the validation procedure (see Sect. 5).

## 5 Applying the Validation Procedure

In this section, we report on the application of the validation procedure (see Sect. 3) to a set of newly generated 4M-models (see Sect. 4). We present the application in the form of an experiment: the current section contains the "experimental method", i.e., gathering behaviour traces in human-in-the-loop simulations (see Sect. 5.1) and performing the assessment (see Sect. 5.2). Later, we present the "experimental results", i.e., the ratings obtained from the assessment and the results of the equivalence tests (see Sect. 6).

### 5.1 Human-in-the-Loop Simulations

Human-in-the-loop simulations were used to determine how a four-ship of red CGFs behaves when the CGFs interact with human participants controlling four blue CGFs. The simulations were performed in NLR's Fighter 4-Ship simulator. This simulator consists of four networked F-16 mock-up cockpits.

The behaviour of the reds was controlled by means of eight 4-models: the four 4P-models plus the four 4M-models (see Sect. 4). Using these eight 4-models, we defined eight *scenarios*. Each scenario was a simulation configuration in which a four-ship of red CGFs approached the human participants from the simulated north. In each scenario, the red four-ship used either one of the four 4P-models or one of the four 4M-models, so that each of the 4-models was used in one of the scenarios.

The human participants in the simulations were active-duty Royal Netherlands Air Force (RNLAF) F-16 pilots from Volkel Airbase (all male, n = 16, age $\mu = 32.0$, $\sigma = 5.35$), and one former RNLAF F-16 pilot (age = 60).[7] No selection

[7] One of the active-duty participants had to leave after four scenarios. This situation presented us with three options: (1) continue without this participant (viz., with a three-ship), (2) cancel the remaining simulations, or (3) substitute the participant with a former F-16 pilot who was available. Since the participant had a non-commanding role in the four-ship, we deemed his influence in the decision-making of the human participants to be minimal. Still, by controlling the fourth blue CGF, he provided valuable input that allowed the red CGFs to function. Furthermore, participants were scarce. We decided that the collection of data was paramount, and let the former F-16 pilot substitute the participant in the remaining simulations.

criteria were applied. The active-duty pilots were assigned to the human-in-the-loop simulations based on availability. Experience levels ranged from *wingman* to *weapons instructor pilot*.

Over the course of three days, five teams of four participants controlled the blue CGFs in the Fighter 4-Ship. Before the simulations took place, the participants received a "mission briefing" document that described (1) the capabilities of the blue CGFs that they would control, and (2) the capabilities of the red CGFs that the participants were to expect in the simulator. The eight scenarios were presented sequentially in a random order. The participants were unaware of the origin of the 4-models controlling the red CGFs (i.e., the simulations were performed in a single-blinded fashion). Each scenario ended when either all four red CGFs, or all four human participants were defeated.

The human-in-the-loop simulations were recorded using the PCDS mission debrief software. In addition to behaviour traces, the recordings included (1) the voice communication that took place among the human participants, and (2) video recordings of the multi-functional displays of the cockpits occupied by the human participants. In total, 33 recordings[8] were stored.

### 5.2 Assessment

The behaviour that the reds displayed in the human-in-the-loop simulations were assessed by human experts. Active-duty RNLAF F-16 pilots from Leeuwarden Airbase acted as assessors (all male, $n = 5$, age $\mu = 35.2$, $\sigma = 5.17$). Assessors were selected on having *tactical instructor pilot* or *weapons instructor pilot* qualification. All five assessors had the weapons instructor pilot qualification. The assessment was performed by means of the ATACC, implemented on paper.

Originally, we had planned to let each assessor assess all of the 33 recordings within a three hour time span. However, a pilot study with two weapons instructor pilots (not counted above) revealed that this was unfeasible because of time constraints. We subsequently reduced the pool of recordings available for rating to 16 recordings. These 16 recordings came from two teams that completed all eight scenarios (i.e., simulations with the four 4P-models and the four 4M-models) in human-in-the-loop simulations. From this reduced pool of recordings, we assigned ten recordings to each assessor, consisting of (1) eight recordings from one of the two teams in random order, and (2) two recordings from the other team. Furthermore, the weapons instructor pilots in the pilot study expressed that they were unable to adequately assess the intelligence of the red CGFs (rating item 8) and the extent to which the red CGFs tested the skills of the pilots in the simulator (rating item 9) without knowing the experience levels of these pilots. Based on this feedback, we made the decision to disclose the experience levels to the assessors during the assessment.

The assessors were provided with a laptop computer with mouse and headphones, a stack of ten ATACC forms, and an instruction sheet. The PCDS recordings were opened on the computer. Each ATACC was marked with a unique code

[8] Two teams were not available to complete all eight scenarios. Together, these two teams completed nine scenarios: the eight scenarios, plus one duplicate.

**Table 1.** Summary of the ATACC responses: the number of responses ($n$), mean response ($\mu$), and standard deviation ($\sigma$) of the responses to the ATACC rating items for the 4P-models and the 4M-models.

| Rating item | 4P-models | | | 4M-models | | |
|---|---|---|---|---|---|---|
| | $n$ | $\mu$ | $\sigma$ | $n$ | $\mu$ | $\sigma$ |
| 1 | 28 | 3.04 | 0.79 | 24 | 3.25 | 0.99 |
| 2 | 28 | 2.07 | 0.98 | 24 | 2.33 | 1.13 |
| 3 | 28 | 3.18 | 1.19 | 24 | 3.92 | 1.02 |
| 4 | 27 | 2.26 | 0.86 | 24 | 2.71 | 0.91 |
| 5 | 28 | 3.29 | 0.71 | 24 | 3.42 | 0.58 |
| 6 | 28 | 2.75 | 0.89 | 24 | 3.33 | 0.70 |
| 7 | 22 | 3.82 | 0.66 | 20 | 3.70 | 0.73 |
| 8 | 28 | 2.86 | 0.80 | 24 | 2.96 | 0.69 |
| 9 | 27 | 3.81 | 0.68 | 24 | 3.63 | 0.65 |

that referred to a specific recording in PCDS. The assessors were instructed to view the recordings in the (pre-randomised) order as indicated by their ATACCs.

## 6 Validation Results

In this section, we present the results of (a) the assessments and (b) the equivalence tests that were performed. Additionally, we provide the results of (c) follow-up tests in the cases where no equivalence was found.

*Assessment Results.* A summary of the responses to the ATACC is given in Table 1. The responses to the Likert scale rating items were coded as integer values ranging from 1 (Never/Strongly disagree) to 5 (Always/Strongly agree). The coding for rating item four (*Blue air was able to fire without threat from red air*) was inverted so that the values reflected the occurrence of red behaviour (i.e., red influencing blue's ability to fire).

*Equivalence Testing.* We applied Schuirmann's (1987) TOST method to determine the equivalence of (1) the responses given on the ATACC for 4P-models, and (2) the responses given on the ATACC for 4M-models. We calculated $\delta$ (as Juzek's $\delta$) for the responses to each rating item of the ATACC, and then performed the TOST on the responses to each rating item. The TOST was performed using the `TOSTtwo.raw` function from R's `TOSTER` package, with Welch's $t$-test as the underlying one-sided test. We chose to use Welch's $t$-test here because of

the unequal sample sizes.[9] The $\delta$ and the results of the TOST ($t$-value, degrees of freedom [$df$], $p$-value, and the 90% confidence interval [CI] of the difference of the means) are shown in Table 2. In Table 2, the bold $p$-values indicate a significant result of the TOST. Based on the results of the TOST, we conclude that the responses to rating items 1, 2, 5, 7, 8, and 9 are equivalent between the 4P-models and the 4M-models (see Sect. 3.2 for the definitions of the examples of behaviour represented by these rating items).

**Table 2.** Results of the TOST method per rating item (i.). The TOST was based on Welch's $t$-test. For rating items where the TOST method did not find equivalence, an additional standard (Welch's) $t$-test was performed. Significant $p$-values at the $\alpha = 0.05$ level are indicated in bold. The relevance (rel.) of the outcome of the tests is indicated in the rightmost column.

| i. | TOST | | | | | Standard $t$-test | | | | Rel. |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\delta$ | $t$ | $df$ | $p$ | 90% CI | $t$ | $df$ | $p$ | 95% CI | |
| 1 | 0.798 | 2.322 | 43.9 | **.012** | −0.637; +0.208 | | | | | eq. |
| 2 | 0.944 | 2.307 | 45.9 | **.013** | −0.758; +0.234 | | | | | eq. |
| 3 | 1.000 | 0.855 | 50.0 | .198 | −1.251; −0.225 | −2.41 | 50.0 | **.020** | −1.353; −0.124 | diff. |
| 4 | 0.800 | 1.414 | 47.5 | .082 | −0.866; −0.032 | −1.81 | 47.5 | .077 | −0.949; +0.050 | und. |
| 5 | 0.590 | 2.551 | 49.9 | **.007** | −0.432; +0.170 | | | | | eq. |
| 6 | 0.725 | 0.643 | 49.7 | .262 | −0.953; −0.214 | −2.64 | 49.7 | **.011** | −1.018; −0.149 | diff. |
| 7 | 0.697 | −2.674 | 38.5 | **.005** | −0.247; +0.483 | | | | | eq. |
| 8 | 0.677 | 2.779 | 50.0 | **.004** | −0.448; +0.246 | | | | | eq. |
| 9 | 0.604 | −2.223 | 48.8 | **.015** | −0.122; +0.502 | | | | | eq |

eq. = equivalent, diff. = different, und. = undecided

*Follow-Up Testing.* The TOST did not find equivalence for rating items 3, 4, and 6. For these rating items, we conducted a follow-up test to determine if the responses to these rating items significantly differed between the 4P-models and the 4M-models. This test was a standard two-sided Welch's $t$-test. A significant difference was found for rating items 3 (*Red air was within factor range*) and 6 (*Red air acted on blue air's weapons engagement zone*). For both rating items, the responses indicated a higher frequency of the behaviour that was rated for the 4M-models (see Table 1). The responses to rating item 4 (*Blue air was able to fire without threat from red air*) were neither significantly equivalent, nor significantly different. Therefore, we may conclude that their relationship is undecided.

[9] There is an ongoing discussion on the topic of whether parametric tests such as the $t$-test are suitable for use on ordinal Likert-scale data. Parametric tests have on multiple occasions been shown to be robust against violated assumptions (such as non-normal, ordinal data) (cf. De Winter 2013; Derrick and White 2017; Norman 2010). Using parametric tests in our TOST allows us to use well-tested, publicly available tools such as the mentioned R package.

## 7 Discussion and Related Work

We started this paper by decomposing the difficulty of validating behaviour models into two questions (see Sect. 2): *what does the process entail?* and *how should we determine the accuracy of the models?* For the case of air combat behaviour models, our answer to the first question is the procedure laid out in Sect. 3. Our answer to the second question is embedded in the procedure: we determine the accuracy of newly generated 4M-models by a combination of simulation technology, behavioural science, statistical methods, and human input.

For our case study, we generated a new set of air combat behaviour models by means of machine learning, and applied the validation procedure to these models. Our key finding is that out of the nine rating items of the ATACC, six are assessed as equivalent between the 4M-models and the 4P-models. Following the advice of Birta and Arbez (2013) to recognise the partial success, the results appear to *moderately* indicate validity. Still, the responses to the remaining three rating items do not support the notion of validity as we have defined it.

Is there any way that we could have achieved a more convincing indication towards the (non-)validity of the new models? We must acknowledge the large number of variables in our study, e.g., (1) the 4P-models, (2) the 4M-models, (3) the pilots, (4) the assessors, and (5) the ATACC. While efforts could be made to control the "noise" from these variable, it is important to consider that (1) and (2) exist in too many variations to ever be sampled effectively, and (3) and (4) are assisting with all the implicit and explicit knowledge they have to offer. The contribution of this knowledge should be stimulated before it is controlled. It is therefore that we propose that improvements should be sought in the area of the assessment tool (5), such as refinement of the examples of behaviour posed by the ATACC. One interesting approach might be to incorporate recent work on the mission essential competencies (MECs) into the tool (see, e.g., MacMillan et al. 2013; Tsifetakis and Kontogiannis 2017).

The validation study performed by Sadagic (2010) most closely resembles our work. The subject of this study were behaviour models for troops in urban warfare. Expert assessors observed the behaviour of these troops, and rated its realism. The work of Sadagic differs from ours in that their simulations had no human participants. Furthermore, no statistical tests were performed, as the behaviour was rated as conforming to the assessors' ideal of realistic behaviour.

In the air combat domain, we find small-scale validation studies attached to machine learning experiments. For instance, Teng et al. (2013) show that their adaptive CGFs are rated more favourably than non-adaptive CGFs on certain qualities (e.g., predictability and aggression) by expert assessors. However, in contrast to our work, Teng, Tan, and Teow aimed to develop CGFs that showed improvement on these qualities, rather than find equivalence. By focusing on improvement, the adaptive capabilities of the CGFs have been validated, but the question remains whether the improved qualities are useful for training.

In conclusion, properly validating air combat behaviour models is difficult to accomplish, yet essential for the training simulations that aim to use them.

The validation procedure that we propose likely is one of many possible solutions. We invite more machine learning researchers and training experts to jointly address the issue of validation in future research, thereby paving the way to reliable adaptive training of teams.

**Acknowledgment.** The author graciously thanks 312, 313, and 322 squadrons of the RNLAF for their generous support during this study. Many thanks to *Rich*, *Gump*, *Slime*, and *Speedy* for sharing their ideas regarding the ATACC and test-driving the simulator. The author is also grateful to Jaap van den Herik, Pieter Spronck and Jan Joris Roessingh for their advice and thorough reviews.

## References

Anderson-Cook, C.M., Borror, C.M.: The difference between *equivalent* and *not different*. Qual. Eng. **28**(3), 249–262 (2016). https://doi.org/10.1080/08982112.2015.1079918

Balci, O.: Validation, verification, and testing techniques throughout the life cycle of a simulation study. Ann. Oper. Res. **53**(1), 121–173 (1994). https://doi.org/10.1109/wsc.1994.717129

Birta, L.G., Arbez, G.: Modelling and Simulation: Exploring Dynamic System Behaviour. Springer, London (2013). https://doi.org/10.1007/978-1-4471-2783-3. ISBN 978-1-4471-2783-3

Bruzzone, A.G., Massei, M.: Simulation-based military training. In: Mittal, S., Durak, U., Ören, T. (eds.) Guide to Simulation-Based Disciplines. SFMA, pp. 315–361. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-61264-5_14. ISBN 978-3-319-61264-5

De Winter, J.C.F.: Using the student's t-test with extremely small sample sizes. Pract. Assess. Res. Eval. **18**(10) (2013)

Derrick, B., White, P.: Comparing two samples from an individual Likert question. Int. J. Math. Stat. **18**(3), 1–13 (2017)

Floyd, M.W., et al.: A goal reasoning agent for controlling UAVs in beyond-visual-range air combat. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, pp. 4714–4721. AAAI Press (2017)

Goerger, S.R., McGinnis, M.L., Darken, R.P.: A validation methodology for human behavior representation models. J. Def. Model. Simul. **2**(1), 39–51 (2005). https://doi.org/10.1177/154851290500200105

Hahn, H.A.: The conundrum of verification and validation of social science-based models Redux. In: Schatz, S., Hoffman, M. (eds.) Advances in Cross-Cultural Decision Making. AISC, vol. 480, pp. 279–292. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-41636-6_23

Juzek, T.S.: Acceptability judgement tasks and grammatical theory. Ph.D. thesis. University of Oxford (2016)

Karli, M., Efe, M.Ö., Sever, H.: Air combat learning from F-16 flight information. In: Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), pp. 1–6 (2017). https://doi.org/10.1109/FUZZ-IEEE.2017.8015615

Karneeb, J., et al.: Distributed discrepancy detection for a goal reasoning agent in beyond-visual-range air combat. In: Roberts, M., et al. (eds.) AI Communications, vol. 31, no. 2, pp. 181–195 (2018). https://doi.org/10.3233/aic-180757

Kim, J.H., et al.: Verification, Validation, and Accreditation (VV&A) considering military and defense characteristics. Ind. Eng. Manag. Syst. **14**(1), 88–93 (2015). https://doi.org/10.7232/iems.2015.14.1.088

Lakens, D.: Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. Soc. Psychol. Pers. Sci. **8**(4), 355–362 (2017). https://doi.org/10.1177/1948550617697177. ISSN 1948-5514

MacMillan, J., et al.: Measuring team performance in complex and dynamic military environments: the SPOTLITE method. Mil. Psychol. **25**, 266 (2013)

Meyners, M.: Equivalence tests - a review. Food Qual. Prefer. **26**(2), 231–245 (2012). https://doi.org/10.1016/j.foodqual.2012.05.003. ISSN 0950-3293

Norman, G.: Likert scales, levels of measurement and the *laws* of statistics. Adv. Health Sci. Educ. **15**(5), 625–632 (2010)

Petty, M.D.: Benefits and consequences of automated learning in computer generated forces systems. Inf. Secur. **12**, 63–74 (2003). https://doi.org/10.11610/isij.1203

Petty, M.D.: Verification, validation, and accreditation. In: Sokolowski, J.A., Banks, C.M. (eds.) Modeling and Simulation Fundamentals: Theoretical Underpinnings and Practical Domains, Chap. 10, pp. 325–372. Wiley, Hoboken (2010). ISBN 978-0-470-48674-0

Sadagic, A.: Validating visual simulation of small unit behavior. In: Proceedings of the 2010 Interservice/Industry Training, Simulation, and Education Conference, I/ITSEC, Orlando, Florida (2010)

Sargent, R.G.: Verification and validation of simulation models. In: Proceedings of the Winter Simulation Conference, WSC 2011, Winter Simulation Conference, Phoenix, Arizona, pp. 183–198 (2011)

Schuirmann, D.J.: A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. J. Pharmacokinet Pharmacodyn. **15**(6), 657–680 (1987). https://doi.org/10.1007/BF01068419

Spronck, P., et al.: Adaptive game AI with dynamic scripting. Mach. Learn. **63**(3), 217–248 (2006). https://doi.org/10.1007/s10994-006-6205-6. ISSN 08856125

Teng, T.-H., Tan, A.-H., Teow, L.-N.: Adaptive computer generated forces for simulator-based training. Expert Syst. Appl. **40**(18), 7341–7353 (2013). https://doi.org/10.1016/j.eswa.2013.07.004

Toubman, A., et al.: Rapid adaptation of air combat behaviour. In: Kaminka, G.A., et al. (eds.) ECAI 2016–22nd European Conference on Artificial Intelligence. Frontiers in Artificial Intelligence and Applications, The Hague, The Netherlands, vol. 285. IOS Press, pp. 1791–1796 (2016). https://doi.org/10.3233/978-1-61499-672-9-1791

Tsifetakis, E., Kontogiannis, T.: Evaluating non-technical skills and mission essential competencies of pilots in military aviation environments. Ergonomics, 1–15 (2017). PMID 28534423. https://doi.org/10.1080/00140139.2017.1332393

US Department of Defense: DoD Modeling and Simulation (M&S) Verification, Validation, and Accreditation (VV&A). Department of Defense Instruction 5000.61 (2009) http://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/500061p.pdf