

## THE IMPORTANCE OF BEING DOCKED

S. K. JOHNSON, The MITRE Corporation, McLean VA  
 M. T. K. KOEHLER\*, The MITRE Corporation, McLean VA  
 D. QUINN, The MITRE Corporation, McLean VA

### ABSTRACT

The concept of docking agent-based models has been stressed in many venues for a number of years. In the decision-support context docking can take on a very important role and we have found docking to be an important exercise not when moving from one framework to another but rather when moving from one version of a model to another. In the decision-support context a great deal of time and energy is spent in the Validation, Verification, and Accreditation cycle so a model may be trusted for a particular use. This paper describes the docking exercise we undertook in order to move confidently from a validated modeling framework to one that had not been face validated. The methodology includes taking very large sample runs from each version of the framework. Output analysis included standard descriptive statistics and non-parametric sample comparisons. Of particular note is that at some levels of aggregation the two models appear to behave very similarly. As one disaggregates groups of agents and looks more closely at the results, however, one begins to find differences. This highlights one of the most important lessons of our docking exercise: just like agent-based modeling, docking must be done at an appropriate level of abstraction for the questions at hand. One must understand the context in which the model(s) are to be used in order to understand what differences are of practical significance and what differences can be tolerated. Our output analysis techniques and results will be discussed in detail.

### INTRODUCTION

*“The purpose of computing is insight, not numbers.”*

~R. W. Hamming

In general, there are a number of reasons why one would want to compare, or dock (Axtell 1996), models. First of all, one simply may be curious about the similarities and differences between two models. These models may seem quite similar or quite different at first glance and a deeper understanding is desirable. Second, one may wish to understand the theoretic difference between models or to understand which model did a better job of representing a common theoretic foundation. Third, one may wish to utilize a model as a decision-support tool and, therefore, must know which model is best and in which contexts. Fourth, one may use the comparison to understand the significance of similarities or differences found in the output of the models. This by no means exhausts the reasons one might wish to compare models but it highlights some of the major ones.

Things change a bit as we move from the academic field to a decision-support context. Here model comparisons become more pragmatic. In many cases the comparisons are done with an eye towards which model is best and in which contexts. This type of comparison can extend

---

\* Corresponding author address: Matthew Koehler, The MITRE Corporation, 7515 Colshire Dr., MailStop H305, McLean, VA 22015, e-mail: mkoehler@mitre.org

to different versions of the *same* model. Furthermore, in the decision-support context it is likely that a model has undergone some sort of verification, validation and accreditation (VV&A). See Koehler, et al. 2006 for a more through discussion of VV&A in this context. Loosely speaking, Verification is determining whether or not you built the model correctly, Validation is determining whether or not you built the correct model, and Accreditation is determining if the model is good enough to use for its intended purpose (Hartley 1997). VV&A can be very time consuming, difficult, and costly. Of course effort put into VV&A is a function of the role the model will play. Models that will be relied upon heavily will under go more rigorous VV&A than will models that are used for thought experiments or to gain rough order of magnitude insights. When moving from one model to another, especially when it is different versions of the same model, any aspects of the original VV&A that can be carried over would be very beneficial. In this particular case we discuss moving from MANA (McIntosh 2007) version 3.0.39 (MANA3) to MANA version 4.00.2 (MANA4). Here docking is used as evidence for the claim that the VV&A done for one model may be valid for another model. We have spent a great deal of time face validating MANA3. Now that we are moving to MANA4 we would like to make the argument that the face validation from MANA3 can carry over (at least to some extent) to MANA4. This may prove difficult, however, as there have been a number of changes made to MANA4. Table 1 highlights some of the more important changes.

**Table 1: Major changes made in MANA4.**

New in MANA4
Agents have orientation and bearing
Agents can move in formations
Agents have sensor and weapon orientation
Agents can have multiple sensors
Sensors can be of multiple types
Sensors and weapons have look angles
Sensors and weapons have slew rates

The evidence we will use for this argument will come from a docking experiment we undertook between MANA3 and MANA4. As discussed by Robert Axtell (Axtell 1996) docking can be achieved in three basic categories. Essentially, docking is the alignment of two different models to understand if one model can subsume another. In the case at hand we wish to say that the two models (MANA3 and MANA4) are not different. Axtell proposes three levels of docking: identity, where two models produce identical results; distributional, where the two models produce statistically indistinguishable results; and relational, where the two models produce output that “behaves” in the same way, meaning that similar changes in inputs cause similar changes in outputs but the distributions are statistically distinct. The necessary level of docking is a function of the empirical relevance of the models, which, in turn, is a function of how the models are relied upon and is the topic of the next section.

### **Specifying the Relevance of an ABM and Level for Docking**

Axtell’s Framework of Empirical Relevance (FER) relates a model to its input data and output (Axtell 2005). The relationship is, generally, input data of various types are necessary to

create an ABM; once the model is run it will create output data that will relate to real-world phenomena in some particular way. There are four levels to Axtell's FER. Level 0 is, essentially, a well functioning program that is bug free. Level 0 models have qualitative correspondence at the agent level. This means that the agents behave in a manner that is logically consistent with the subject being modeled. Level 1 is the next level of Axtell's FER. Level 1 is macro-level qualitative correspondence to the dynamics of interest. In this level the agent activity generates dynamics, as a whole, that relate to the phenomena being modeled. For example, a group of agents trading with each other may produce a clearing price for the artificial market. This clearing price may not relate to a real-world clearing price but one was found. Axtell's levels continue with Level 2. Models that fall into this category have macro-level quantitative correspondence with the real-world phenomena being modeled. These models produce the correct distributions within their output. For example, Axtell's model of firm size produces a power-law distribution that is the same as real-world data on the distribution of firm sizes in the US (Axtell 2001). The final level of this framework is that of Level 3. In this level the model not only has macro-level quantitative correspondence but also micro-level correspondence. In general, very few ABMs achieve this level of empirical relevance. This is because it is difficult to specify such a model and even more difficult to obtain the data necessary to estimate such a specified model.

In the decision-support context the weight put upon model output and the importance of decisions based upon said output will necessitate that the model achieve a particular level of the FER. Here MANA is being used in a decision-support context to aid subject matter experts (SME). MANA is only one part of the whole decision-support infrastructure. Furthermore, MANA is to be very fast turnaround. Therefore, MANA is understood to provide rough order of magnitude answers to what-if analyses. This places MANA squarely on Level 1 of the FER scale. MANA needs to be in qualitative agreement at the macro-level and have reasonable micro-level behaviors. This implies that identity between MANA3 and MANA4 is not necessary to conclude that they are equivalent. Therefore, distributional equivalence is adequate to conclude that these models are equivalent, and relational equivalence may be adequate.

## RESULTS

As we have already stated, we are interested in showing distributional equivalence of the two versions of the simulation, rather than identity. In our docking experiment, only the versions of MANA were different. We used the same scenarios with the same random seeds and ran the same number of sample runs through the two versions of the simulation.

We tested several types of scenarios, with variable levels of complexity in terms of terrain, agent behaviors, agent interactions with each other and their environments, communications, and weapon systems. Our objective was to test a set of scenarios with low, medium, and high complexity in both versions. We hypothesized that our results would show statistically indistinguishable results between versions since we did nothing to the scenarios other than port them into each model and run them a large number of times. However, if significant differences occurred, would they happen across all of the scenarios, to include the simplest set of scenarios, or would they happen in just the more complicated cases?

The low level complexity scenarios simply exercised important features of the models in the simplest manner possible. The medium level complexity scenario added slightly more complicated terrain, weapons, communications, agent and squad behaviors, and contained a few

more squads than the low level complexity scenarios. The high level complexity scenarios have the same terrain as the medium level complexity scenario; however they have a much more sophisticated communications structure, more types of agents, a much larger variety of more complicated weapons, as well as more complicated agent interactions. We setup our experiment to test a variety of scenarios to determine if VV&A should be done on each scenario moved from MANA3 to MANA4 or if we could assume based upon this docking experiment that VV&A is independent of the MANA version and would transfer from MANA3 to MANA4 with the scenario.

Though MANA produces a reasonable set of output data, for the sake of simplicity we chose to use the most basic MANA output statistic: casualties. First, we did a crude comparison of raw casualty numbers by scenario, by replicate, and by side between the two versions of the simulation and found that identity did not exist for any of the scenarios. However, since we do not require identity, the next set of tests explored whether or not the differences between the versions of MANA were statistically significant or if we could claim distributional equivalence of the versions.

We treated the data as a paired sample for our experiment because we used the same scenario with the same random seeds with the before and after elements represented by the two versions of the simulation. We ran the parametric Paired Sample t-Test and the nonparametric Wilcoxon Signed Rank Test and Sign Test as our test procedures. The Paired Sample t-Test assumes the data comes from a normal distribution. The Wilcoxon Signed Rank Test does not assume the distributions of data are normal, however it does assume the distributions are symmetric. The Sign Test does not assume normality, nor does it assume symmetry, but it has less power to detect significance if there is symmetry. Review of the Q-Q plots for each of the datasets indicated that most of the distributions had heavy tails, and therefore appeared to deviate from a normal distribution. Evaluation of the measurement for skew, which is one indicator of how far from symmetric distributions are, clustered the majority of values for each of the paired differences in the scenarios tested within the -0.1 to +0.1 range. Since most of the data sets appeared inconsistent with a normal distribution and since their values for skew did not tend to be too extreme, we are reporting the Wilcoxon Signed Rank Test results in this paper. Of note however, in almost every case, the three tests produce the same results with respect to statistical significance (Bhattacharyya 1977; Gibbons 2003).

**Table 2: Wilcoxon Signed Rank Test results between MANA3 and MANA4 for the highest level of aggregation for each scenario complexity level.**

Complexity Level - Scenario Name	Pairs	Wilcoxon Signed Rank Test
Low - Four Groups	Blue Casualties MANA V3 to MANA V4	0.279
	Red Casualties MANA V3 to MANA V4	0.141
Low - Two Groups	Blue Casualties MANA V3 to MANA V4	0.143
	Red Casualties MANA V3 to MANA V4	0.311
Low - Two Groups with Communications	Blue Casualties MANA V3 to MANA V4	0.138
	Red Casualties MANA V3 to MANA V4	0.980
Medium - Scenario 1	Blue Casualties MANA V3 to MANA V4	0.797

	Red Casualties MANA V3 to MANA V4	0.000
	Civilian Casualties MANA V3 to MANA V4	0.259
High - Scenario 1	Blue Casualties MANA V3 to MANA V4	0.000
	Red Casualties MANA V3 to MANA V4	0.000
High - Scenario 2	Blue Casualties MANA V3 to MANA V4	0.000
	Red Casualties MANA V3 to MANA V4	0.000

In addition to testing several levels of scenario complexity, our experiment also consisted of testing the same metric at various levels of data aggregation for the low, medium, and high complexity sets of scenarios. The highest level of aggregation is represented by the Blue, Red, and Civilian casualty pairs. Table 2 shows the p-values for each of the pairs tested for each scenario between the two versions at the highest level of aggregation. None of the casualty pairs tested for the low level complexity scenarios had significant results (all p-values  $> 0.1$ ), whereas all of the casualty pairs tested for the high level complexity scenarios had statistically significant results (all p-values  $< 0.0$ ). The medium level complexity scenario returned significance only for the Red casualty pairs between the two versions (Red p-value  $< 0.0$ ), but the Blue and Civilian comparisons were not significant (Blue p-value = 0.797; Civilian p-value = 0.259). At this level, what these results imply is that we cannot claim distributional equivalence across the board, but we can make the claim for distributional equivalence in some instances. For our experiment, this means that the simple scenarios are distributional equivalent. However, this does not hold once we increase the level of complexity. Unfortunately, distributional equivalence appears to be scenario dependent.

However, is the presence or absence of distributional equivalence only scenario dependent or does it also depend on the level of data aggregation? Do the pairs that exhibit non-significance at the highest level of aggregation also show non-significance for each of their respective sub-categories when the datasets are disaggregated? The next step to the experiment was to break the data into sub-categories and run the same statistical tests using the same metric, total casualties by sub-category by side. This process was repeated for each of the scenarios. We did this to see if we could determine the largest contributors to the differences. For space reasons only the data for the medium level complexity scenario is reported.

We hypothesized that we would see the same outcomes when we disaggregated the data into sub-categories as we did for the highest level of aggregation. For the low complexity scenarios, all of the sub-categories by pair had non-significant results. As expected, in the high complexity scenario nearly all of the Blue and Red pairs returned significant results. Interestingly, however, in the medium complexity scenario we found that we did not have distributional equivalence when comparing the sub-categories. Tables 3 and 4 display the p-value results for the medium complexity scenario disaggregated into killer and victim squad categories respectively. The data in these tables appear to support the idea that distributional equivalence is dependent not only on the scenario and the complexity of the scenario, but also on the level of data aggregation.

In Table 2, for the medium complexity scenario we saw that the Blue casualty and Civilian casualty pairs did not have statistically significant results between the two model versions, whereas the Red casualty pairs were significant. Tables 3 and 4 show mixed results for the individual squad pairs for both Blue and Red. In Table 3, half of the Blue squads had significant results and half did not and almost all of the Red pairs returned non-significant

results. This data indicates that between the two versions of MANA the Red killer squads were the same and had about the same number of casualties attributed to them; however, the Blue killer squads were split and did not necessarily have the same number of casualties. If we look at Table 4, we see that for the majority of the Blue and Red pairs, we also get significant results, which suggests that the victim squads were not necessarily the same between the two versions. The results for Red are reasonable because the results at the highest aggregation level for Red were significant. However, the results for Blue in Table 4 seem counterintuitive. At the highest level of aggregation, the results for the Blue casualty pairs were not significant between the two versions, but as we disaggregated from the highest level, the results became significant. These outcomes indicate that the victim squads for Blue were not necessarily the same between the two versions, even though the overall number of casualties for Blue was not significantly different.

**Table 3: Wilcoxon Signed Rank Test results between MANA3 and MANA4 by killer squad category for medium complexity level scenario.**

Complexity Level - Scenario Name	Pairs	Wilcoxon Signed Rank Test
Medium - Scenario 1	Blue Kills on Red	.
	Blue Advance Guard	0.000
	Blue Cargo Truck	0.360
	Blue Convoy Commander	0.027
	Blue Forward Security	0.525
	Blue Fwd Security Command	0.004
	Blue Gun Truck	0.837
	Blue Rear Security Command	0.000
	Blue Rear Security	0.734
	Red Kills on Blue	
	Red IED 1	0.259
	Red IED 2	0.285
	Red Attack Vehicle	0.000
	Red IED 3	0.157
	Red RPG	0.250
	Blue Kills on Civilians	
	Blue Advance Guard	0.518
	Blue Cargo Truck	0.892
	Blue Convoy Commander	0.063
	Blue Forward Security	0.039
	Blue Fwd Security Command	0.152
	Blue Gun Truck	0.148
	Red Kills on Civilians	
	Red RPG	0.004

**Table 4: Wilcoxon Signed Rank Test results between MANA V3 and MANA V4 by victim squad category for medium complexity level scenario.**

Complexity Level - Scenario Name	Pairs	Wilcoxon Signed Rank Test
Medium - Scenario 1	Blue Victims	.
	Blue Advance Guard	0.285
	Blue Cargo Truck	0.000
	Blue Convoy Commander	0.000
	Blue Forward Security	0.000
	Blue Forward Security Command	0.000
	Blue Gun Truck	0.000
	Red Victims	
	Red Pickup Truck	0.000
	Red IED 2	0.157
	Red Attack Vehicle	0.000
	Red Rifle Squad	0.000
	Red RPG	0.006
	Red Sniper	0.227
	Red Observer	0.044
	Civilian Victims	
	Civilians	0.259

## CONCLUSION

As highlighted above docking can be a difficult undertaking even when comparing seemingly equivalent models. We set out to determine if MANA3 was distributional equivalent to MANA4. Clearly, except in trivially simple cases we cannot make that claim. Where does that leave us? Ultimately it is up to the decision-maker to decide if any of the statistically significant results have practical significance within the decision-support context in question. If the models function in a sound manner, and if the distributions of data between versions, even though statistically significant, differ by a small amount, then the decision-maker may still accept the models as equivalent. This is the case because the statistical significance may have little or no practical significance. However, one cannot, *prima impressionis*, claim that a VV&A assessment will move from MANA3 to MANA4.

It should be noted, however, that this comes as no surprise given the numerous changes highlighted in Table 1. The most significant of these changes include agent orientation and the behavior of sensors. In MANA3 agents and sensors had no orientation; there was no difference between front and back. In MANA4 agents and sensors have, *inter alia*, an orientation and a speed associated with changing orientation. Agent success and failure is premised highly upon an awareness of the environment. Therefore, changes to the behavior of sensors and the way

agents “look around” the environment will impact scenario results. MANA4’s inclusion of orientation for both agents and sensors significantly increases the verisimilitude of the framework. Finally, as we wanted to make as direct a comparison as possible between the two versions of MANA *no attempt* was made to change default settings in MANA4 to better emulate behaviors of MANA3. Given the higher verisimilitude of MANA4 one, in fact, may not want MANA4 to exactly match MANA3.

**Table 5: Medium complexity scenario descriptive statistics for MANA3 (Old) compared to MANA4 (New)**

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
BlueOld	500	4.016	1.68089107	1	15	3	4	5
BlueNew	500	3.972	1.71262435	2	14	3	4	4
RedOld	500	16.796	2.264650367	7	22	15	17	19
RedNew	500	16.16	2.17896668	9	21	15	17	18
CivilianOld	500	10.404	3.25052547	2	19	8	11	13
CivilianNew	500	10.162	3.066359588	1	20	8	10	12

**Table 6: High complexity scenario descriptive statistics for MANA3 (Old) compared to MANA4 (New)**

	N	Mean	Std. Deviation	Minimum	Maximum	Percentiles		
						25th	50th (Median)	75th
BlueOld	394	138.8654822	13.48921637	57	180	134	140	146
BlueNew	394	128.1395939	3.685237151	116	138	126	128	131
RedOld	394	259.1142132	3.213307415	248	267	257	259	262
RedNew	394	270.5025381	2.999786872	261	279	268.75	271	272

This, again, highlights the role of the subject matter expert and decision-maker. For example, if we recall from Table 2, the results for Red were statistically significant. But if we look at Table 5 the Red means between the two versions (MANA3 = Old; MANA4 = New) do not differ by a large factor and the distributions of data between the two versions are similar. In this instance, a decision-maker may conclude that the results are not practically significant and that the models are equivalent. However, this may not be the case for every scenario. Recall that in Table 2, the results were statistically significant across the board for the high complexity scenarios. The descriptive statistics in Table 6 for one of the high complexity scenarios indicate that the distributions of data for the Blue pairs differ by a much larger factor. In this case, the decision-maker may deem the results practically significant, because in reality, this may mean the difference in whether or not an operation is halted or continued. In this case, carrying over the VV&A of the scenario may not be possible; thus, necessitating a new VV&A cycle for MANA4.

**REFERENCES**

- Axtell R., 2005, "Three Distinct Kinds of Empirically-Relevant Agent-Based Models," Brookings Institution Center on Social and Economic Dynamics Working Papers.
- Axtell R., 2001, "Zipf Distribution of U.S. Firm Sizes," *Science*, 293(5536), 1818-20.
- Axtell R, R. Axelrod, J. Epstein, and M. Cohen, 1996, "Aligning Simulation Models: A Case Study and Results," *Computational and Mathematical Organization Theory*, 1:123-141.
- Bhattacharyya, G. and R. Johnson, 1977, *Statistical Concepts and Methods*, John Wiley & Sons, Inc.
- Gibbons, J., and S. Chakraborti, 2003, *Nonparametric Statistical Inference*, Marcel Dekker, Inc.
- Hartley, D., 1997, "Verification & Validation in Military Simulations," in *Proceeding of the 1997 Winter Simulation Conference*, K. Healy, D. Withers, and B. Nelson, editors.
- Koehler, M, P. Barry, and T. Meyer, 2006, "Sending Agents to War," in *Proceedings of the Agent 2006 Conference*. Chicago, Il: Argonne National Lab.
- McIntosh, G., D. Galligan, M. Anderson, and M. Lauren, 2007, "Recent Developments in the MANA Agent-based Model," in *The Scythe*, issue 1, The Naval Postgraduate School, Monterey, CA.

