Generalizing a Numeric Personality Metric for Narrative Planners

Elinor Rubin-McGregor, Brent Harrison

[1]Department of Computer Science University of Kentucky, Davis Marksbury Building, 329 Rose Street, Lexington, KY 40506-0633 USA

Abstract

In the field of narrative planning, there are many different approaches to personality modeling. So many that overarching study of personality models themselves is beginning to form. But a subject as complex as personality demands complex modeling, which in turn makes it difficult to compare implementations or to test sub-features of personality systems intended to be globalized. By generalizing an existing five-number model personality system, we hope to provide an adaptable resource that can be used for enhancement, comparison, or simply providing a foundational basis to other personality models.

1. Introduction

The consideration of personality is a major step forward in the field of narrative generation. Narrative generators have a multitude of applications, from training models to video games and even organizational and strategic purposes. Incorporating personality into narrative models is a subject that has vexed many researchers for years, as personality is such a complex and varied concept. Yet it is critical, for if we do not model personality, our narratives cannot consider ways in which behavior differs between different individuals. For narrative purposes alone, stories become more engaging if the audience can identify with the characters and see them as reflections of real people. Without personality considered, it is far more difficult to display narrative elements known to entice audiences such as character depth. Two people in the same situation will make different choices depending on who they are, and attempting to capture that concept of "who they are" has been the pursuit of many.

Currently, there are a wide variety of different unique personality models proposed for this purpose, with varying advantages and disadvantages. Many of these models, however, require a great deal of effort to implement because they rely on information that is difficult for narrative planners to collect. For systems where personality is the central focus, or where personality is an important element this may be an acceptable cost to pay. But what about when the program is not focused on developing a specific personality system, but instead on features related to multiple personality adaptation systems? Or perhaps, when personality is required or beneficial but not the primary focus [1, 2]? What about simply having a baseline personality model to compare a more complex model to [3]? Having a small-scale easily implementable personality model would be beneficial for other researchers in this field.

There is an existing personality model that does not require a great deal of effort to collect, running on data that many narrative planners can easily collect already. This is the OCEAN-based personality model produced by Shirvani and Ware 2019. For easy of understanding and brevity, we refer to this model as Shirvani19. While this model does utilize data that is generally available to narrative planners, it does have limitations associated with it. For example, this model is not entirely open to all domains and has some features that cannot be calculated by a computer during story generation.

Specifically, the Shirvani19 model uses metrics that demand an understanding of not only the current story plan, but all or a large number of hypothetical alternative story plans. One such metric describes "creative thinking." This metric is calculated by checking how many times the specific actions of a given character occur in a larger, preferably all-alternative-plan-encompassing, set of alternative stories. In addition, the paper uses the concept of "conflict" in its metrics for determining both agreeableness and intellect but defines its measurement of conflict as any time a character can observe any way in which their plans can fail. This feature also requires knowledge that cannot easily be generated during story creation, as evaluating it requires essentially finishing the story in multiple ways before the story is even concluded. In short, there are features of the Shirvani19 model that can only be used to evaluate personality after several stories have already been generated, which in turn makes the model difficult to use if we want to apply it during story generation.

To this end, we are proposing to modify the Shirvani19 model such that it can be applied to a wider variety of narrative planners. We are also trying to simplify the overhead required to make the personality model work. Specifically, we propose to calculate a metric that describes "creative thinking" by comparing the diversity of actions only along the specific plan, so that characters who utilize a broader range of actions are considered to have a higher Openness score than characters who repeatedly use the same actions. Likewise conflict is redefined for both of its uses. Where it is applied for measuring a character's affability, we instead check simply the number of ways a character's actions could directly harm other characters. Where conflict is applied to intellect, we translate the chance of success to the chance that other characters will oppose the actions of the given character.

In order to ensure that our proposed methods are usable, we performed a user study wherein subjects evaluated the stories produced by our modified model. In the end we found that while our Agreeableness work seems to be very applicable, our re-definition of Openness will need some refinement in later work.

2. Related Work

There is a large amount of work on representing personality in digital narratives, even work that focuses on the Big Five OCEAN framework, but not many that are very modular [5, 6]. Shirvani came out with a follow-up to Shirvani19 that addressed the issues discussed here, but at the cost of increasing the size of the model [7].

The well-known *Versu* drama manager is very good at

AIIDE Workshop on Intelligent Narrative Technologies, November 18, 2024, University of Kentucky Lexington, KY, USA

erru227@uky.edu (E. Rubin-McGregor); bha286@g.uky.edu (B. Harrison)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

telling complex stories with consistent character personality, but it requires a great deal of overhead work to run [8]. The model needs files representing the world, social practices, and the characters. Not only that but it needs parser programs for all of these features, then initialization functions, then a database to hold it all, and multiple levels of instantiators before it can make a decision. Likewise, the *Comme il Faut* project handles complex emotional environments very well, but it requires a great deal of information provided to the data manager for any story domain to work[9]. Information on cultural knowledge, social facts, social states, social exchanges and even more must be documented in order for them to be applied.

There are of course works that focus less on societal impacts as a whole, and more on the individual characters. Bahamon and Young introduced other OCEAN-based systems, but they have not produced a way to directly evaluate the OCEAN traits during runtime without extensive preparation. Their earlier work in 2012 provides a way to remove actions deemed out-of-character during story generation, but does not provide a mechanism to determine whether behavior is out-of-character or not. It is a model we would like to use to test our own work on in the future [10]. Their later work further discusses evaluating personality consistency in narrative models, but still does not introduce a personality model to use [11].

The drama manager from *Why Are We Like This* works well with the player's actions and models character personality from player actions, but because of this it only works for the specific high degree of player interaction used in the project [12]. It also uses abstract personality modeling, rather than a personality system that can work as soon as it is applied. There has also been work by Soares that models the personality of the player for narrative decisions, but it does not model the characters of the narrative in the same way [5]. Shirvani and Ware developed a very impressive emotion-based personality model that solves many of the same problems this paper seeks to correct [7]. This model relies upon its emotional system heavily in modeling personality, which in turn requires a larger amount of overhead and thus isn't as modular as this paper seeks to be. Its reliance on its emotional system also prevents it from being used with various other emotion-focused models [13, 14].

3. Background on the Shirvani19 Model

Shirvani and Ware proposed a personality model [4] for characters in a computational narrative that was based on the OCEAN model of personality, using Sabre as the basis of its planning model [15]. The OCEAN personality model, or "the big Five model," utilizes five key attributes to collectively describe personality: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism [16]. These are commonly accepted attributes of personality, and are defined as such:

- Openness means "openness to experience" and describes how much a person is willing to explore outside of their comfort zone. This feature is also considered an aspect of curiosity, and therefore is often tied to creativity as well
- Conscientiousness is how organized and effective a person is. Someone who acts carelessly or struggles to complete tasks is considered to have low conscientiousness.
- Extraversion is the degree to which a person wants to engage and interact with other people. Notably a highly extraverted person can also be very malicious, as this category does not differentiate between positive or negative engagement with others, only frequency.
- Agreeableness is reflective of compassion and empathy, and is used to measure how much a person considers other people. Like Extraversion, someone can be very shy and have high agreeableness.
- Neuroticism is a more internal emotional feature, as it describes essentially how nervous and insecure a person is. Highly neurotic people will often struggle with self-esteem, and emotional instability is often linked to high levels of neuroticism.

The Shirvani19 model is primarily focused on scoring the actions of a character according to how those actions relate to these attributes. That is to say, it estimates what personality traits are being displayed in a given character's actions, and to what degree each action displays those traits.

They do this by calculating twelve variables that are each used to contribute to a score describing a different OCEAN attribute. A full table of these metrics and how they relate to each OCEAN attribute are listed in Table 1. Of note, any value with a (R) in it means that the value is used to reduce the overall score, as it defines a facet that makes an action fit less into the given personality attribute.

Of these scores, two are not as easy to formulate as others. Agreeableness and Openness utilize metrics that are difficult to obtain during story generation. We will discuss these metrics in greater detail below.

3.1. Agreeableness

As can be seen in Table 1, the Agreeableness OCEAN quality contains 4 metrics associated with it. One of these metrics, (11 in Table 1), requires the planner to be able to calculate the number of conflicts created for other characters. The Shirvani19 definition for conflict can be problematic for efficient calculation. Shirvani19 defines character conflict as occurring when a character can foresee any way their plan can go wrong and fail to reach their goal. This element is extremely difficult to evaluate in many systems, as it requires simulating all possible alternative actions or events that could happen, not simply the actions they intend to have happen. This would take a large amount of operational time and resources to run, as well as require limits or ways to determine when to stop simulating additional possible future plans.

3.2. Openness

The Openness attribute of OCEAN is defined by Shirvani with two metrics. We refer to the first metric as "creative thinking" (referred to as the openness facet in Table 1) and the second as intellect (1 and 2 in Table 1, respectively). Creative thinking is a variance value, as it is used to reward using a diverse set of actions. The equation for creative thinking is as follows:

$$\text{Creative Thinking} = 1 - \underset{i=1...m}{min} \sum_{j=1}^{n} \frac{Occurences(a_i,p_j)}{Length(p_i)}$$

OCEAN Quality	Facet	Description
Openness	Openness	1.The minimum action likelihood in a plan (R)
	Intellect	2.Probability of success of a plan
Conscientiousness	Industriousness and Orderliness	3.# of actions in a plan (R)
		4.# of times the agent changes their mind (R)
		5.# of actions with self as the consenting character
Extraversion	Enthusiasm	6.# of actions including others with their consent
	Assertiveness	7.# of actions including others without their consent
Agreeableness	Compassion	8.# of actions including others with their consent
		9.# of goals achieved for other characters
	Politeness	10.# of actions including others without their consent (R)
		11.# of conflicts created for other characters (R)
Neuroticism	Withdrawal and Volatility	12.# of times the agent changes their mind

Table 1
Shirvani19's Metrics of Personality for the OCEAN personality model [4].

In this function, we assume the agent is considering n possible different plans to take. The set of these plans is $[p_1 ... p_n]$, so p_i is the i-th plan being considered. The value a_i is a given action in one or more of these plans. Thus we can think of the plans as sets of actions, $p_i = [a_1 ... a_m]$. The value m is the total number of actions that are possible for the character to take. As for the larger values, Occurences(a_i, p_j) is used as the number of times action a_i occurs in plan p_j, while Length(p_i) is the number of steps in plan p_i.

The second metric that contributes to Openness is Intellect. The Shirvani19 model defines this metric as the probability that a plan succeeds. The probability of success of a plan is defined as the likelihood of a plan succeeding based on the number of conflicts created with other characters. This was defined as such:

$$\text{Probability of Success} = 1 - \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{Conflict_{a_j}(c_i)}{n \cdot m}$$

In this, the values of n and m represent the total number of characters and the total number of possible actions respectively. The value c_i represents character i out of the set of all characters in the domain, and a_j represents a given action a_j in the set of all m potential actions. $Conflict_{a_j}(c_i)$ is therefore a value that is 1 if action a_j causes a a conflict for character c_i, and 0 if it does not. In short, the probability of success is defined by how many other characters would agree with the given character's action plan.

Both the Creative Thinking metric and the Intellect metric share some issues in how they are calculated. In both cases, the set of n plans $[p_1 ... p_n]$ demands that the program collect a large collection of potential actions for every single character's potential plans. This works on the assumption that the implementation of personality is done after the planner has generated multiple plans, and assumes that personality is simply used to collect the best possible plan. Such a system is not feasible if the planner is intended to be used for real-time story generation, or if the planner is working with a human agent. It demands not only a large portion of work be completed multiple times for every character on every step, it also needs to have all or a large set number of solutions generated for the metric to be collected.

The Intellect metric is also problematic in that the probability of success calculation relies on being able to calculate whether an action would generate a conflict with another character. We have already discussed the potential issues with calculating conflict information in the previous section.

4. Methods

To make the personality model more flexible, we replaced the problematic aspects of the Openness and Agreeableness OCEAN metrics with values that could be collected more easily. For Agreeableness we only needed to re-evaluate the concept of conflict, but for Openness we propose alternative calculations for both Creative Thinking and Intellect. We will discuss each of these in greater detail below.

4.1. Conflict

In the Shirvani19 model, conflicts are calculated by determining any point at which their plan could fail. While this is a rigorous way to determine conflict, we propose a metric that relaxes the idea of conflict in the interest of making it easier to calculate. Instead of calculating conflict so directly, we propose defining character conflict by the character's goals or other motivating factors. Rather than simulating an entire world change for potential issues, we argue that simply checking two states for comparison is enough. Specifically, our metric compares one existing state and one hypothetical state. The "true" state, t_0 is the state at the moment when the character is considering a plan, before taking or deciding on an action, and is thus "true" because it has come to pass outside of the character's plans. The hypothetical state is the predicted end state that will come to pass if the character's entire plan is executed without fail, t_n. In this we consider t_1 to be the first action in the plan the character is considering, with the considered plan having a total of n steps in it.

Thus, our changed definition of $Conflict_{a_j}(c_i)$ will need to specify that a_j would result in the world state t_j if executed. With this, $Conflict_{a_j}(c_i)$ is 1 if character c_i has a higher goal metric at t_0 than at t_j, and is 0 otherwise. In other words, as long as action a_j moves the character, c_i further from its goal, then $Conflict_{a_j}(c_i)$ will evaluate to 1.

4.2. Creative Thinking

Recall that to calculate Creative Thinking, the Shirvani19 model needs to calculate the variance associated with a plan by calculating the minimum action likelihood in a plan across many different plans being considered. To make creative thinking easier to calculate, we propose to simply examine the diversity of the actions considered in the plan. We define $plannedSet$ as the combined set of all actions that have occurred up to the point at which an action is being considered combined with the set of actions in the most likely future plan. We then define $actionCounter$ as a set of size m, with m being the total number of possible actions in the domain. The set $actionCounter$ holds the number of times every given action in the domain is executed throughout the entirety of the $plannedSet$, and thus can be calculated by going over the $plannedSet$ just once. In other words, $actionCounter$ is the eventual count of how many times every potential action would occur if the given plan occurs without any interference or changes. We calculated a variance-based metric that scales from 0 to 1. Using an existing commonly used variance algorithm, we applied a variance-based metric. In other words, we measured openness to new experiences as the variance between different kinds of action the character showed.

4.3. Intellect

Originally, Intellect was calculated as the probability of success in a plan in terms of the number of conflicts that it could create. We decided the easiest way to simplify the problem is to cut out a unique metric entirely, and instead use the “politeness" metric (10 in Table 1) for two values and two purposes.

Politeness is a metric that is calculated by determining the number of actions that include other characters without them consenting to the action. The first way we utilize the Politeness metric is in its originally intended way. That is to say, it is used to help calculate the Agreeableness of an agent where the smaller the number of actions taken that include nonconsenting characters, the larger the Politeness metric.

The new way we propose to use the Politeness metric is to apply it to the concept of "Opposing Forces" in the sense that it estimates how much opposition the character would need to overcome to ensure the plan operates smoothly. In other words, we consider the plan more likely to succeed- and therefore more intelligent- based on the number of agents that would oppose the plan by not consenting to take part in certain actions.

5. Experiments

To evaluate the quality of our proposed metrics, we ran a human subjects experiment. We attempted to run our experiments as closely as possible to the experiments performed in [4], thus our experiments were focused on whether or not an audience reading our generated stories observed the intended personality traits assigned to a given character. Since our metrics only affect 2 of the OCEAN traits, we limit our experiments to the following four basic treatments involving the OCEAN traits Openness and Agreeableness: High Openness (HO), Low Openness (LO), High Agreeableness (HA), and Low Agreeableness (LA). The experiment was set up as a between-subjects experiment where participants were randomly sorted into one of these four groups.

5.1. Story Domain and Story Generation

Shirvani and Ware unfortunately did not keep track of their original program, thus we were unable to use the exact same domain as they did. For our story experiments, we emulated the domain used in the 2019 work as closely as possible. Thus, we generated stories about a boy named Tom, whose goal is to gain some herbs for his sick grandmother. The herbs are in the possession of a Merchant, whose goal is to gain a coin- which Tom happens to have. The Merchant is in the Town, while Tom is in the Forest. But there is also a Bandit in the Forest, who also wants the coin. Any character can walk from one location to another, any character can buy an item from another by spending a coin, and any character that holds a weapon can rob another character for any item in their inventory. A character with a weapon can also kill another character, and any living character can loot the corpse of a dead character for any items they hold. There is also a bandit camp, where there is a chest with a secondary coin. Finally, there is a guard in Town, who has the unique action to arrest the bandit, and whose goal is to arrest the bandit. The three characters that hold weapons at the start of the story are the bandit, the merchant, and the guard.

For each of the treatments mentioned above we needed to generate a total of four stories. One of these stories was the “true” story. In these stories, Tom would take actions that either ranked very highly in the Agreeableness or Openness metrics as described above, or very lowly in those metrics, depending on the category.

Another story involved Tom displaying the opposite personality to the one being tested. Thus, if the treatment group was associated with Low Agreeableness, then this story would involve Tom performing High Agreeableness actions. This story is meant to be a bad fit for the character’s "true" personality.

The final two stories consisted of a story that had a medium score in the given attribute being tested according to our metrics (and, thus, did not display Tom as strong exhibiting or not exhibiting the attribute in question) and a story chosen at random.

5.2. Experimental Methodology

We had two hypotheses that we wanted to evaluate for our experiments. The first was that when shown a story that our model generated and claimed displays one of our target personality traits in the main character, the audience will identify the main character as someone who holds these traits. The second hypothesis is that when shown a set of stories that includes one tale that our model claimed also displays the same target trait in a similar quantity in the main character, the audience will identify that particular story as the one they consider most realistic.

For these experiments, subjects were first given a brief description of the domain, introducing the people, the places, and the goals of the characters. They were then shown the four generated stories described above and told that all were possible ways that the story could proceed. One of these stories would be the "true" story wherein the target character’s behavior closely matched the personality the subject’s category was testing for. The domains for these stories were

OCEAN Quality	Question
Agreeableness	Tom avoids conflict.
Agreeableness (R)	Tom takes advantage of others.
Agreeableness (R)	Tom is out for his own personal gain, with his grandmother as the only exception.
Agreeableness	Tom likes to do things for others as well as his grandmother.
Agreeableness (R)	Tom can't be bothered with other's needs (unless they are his grandmother).
Extraversion	Tom feels comfortable around people.
Neuroticism	Tom does things that he later regrets.
Conscientiousness	Tom makes plans and sticks to them.

Table 2
Statements Evaluated for testing Agreeableness

identical. In total, there were eight stories shown to subjects testing Openness and eight different stories shown to subjects testing Agreeableness.

After reading this, they were then told that one of the stories was the "true" story and were then asked to rate statements about the target character using a 5-point Likert scale. We tried to use the same statements as Shirvani and Ware, however preliminary testing showed that the framing device for the story was interfering with the results. To be specific, the domain in which the story takes place features the target character of Tom, trying to get herbs for his sick grandmother. We attempted to mention in the explanation that Tom's grandmother provides for him, thus implying a potential selfish motive for Tom's behavior, but the majority of results in our initial testing showed high Agreeableness regardless of Tom's behavior in the story. Thus, we modified the statements slightly so that the participants would give answers based only on the parts of the story that our model had generated rather than the backstory. The statements we used are shown in Table 2. It should be noted that there is no "grandmother" character included in any story domain used, as the character is a plot device and cannot take any actions during the story.

While most statements presented to the user were related to the specific OCEAN category we were testing, we included a few statements related to different OCEAN qualities as well. This was done to avoid having the subjects fixate too heavily on the general theme behind the questions and to encourage them to think about the entire story in their responses. These statements were not used for analyzing the target metric of the category. The statements tested for both HO and LO were the same, as were the statements tested for HA and LA. Table 2 contains the statements presented to the subjects for Agreeableness tests, and Table 3 contains the statements presented to the subjects for Openness tests. It should be noted that some questions were meant to reflect a low score in the given metric, not a high one. Ones marked with an (R) for "Reverse" were expected to be agreed with if the "true" story ranked the character as having a low value in the given metric.

After rating these statements, subjects were then shown four more stories and asked which one they thought most likely to occur based on the target character's personality. In order to increase variability in stories, the second set of stories were generated from two domains that were almost entirely identical to the original domain. These two domains had exactly one additional feature each; one added the location of a Bandit's Camp where a coin could be found, and the other included an additional Guard character whose goal was to arrest the Bandit. The subjects were told that these additional four stories included the bandit camp/the guard, so they would still understand the limitations of the world. No other changes were made to the domains. Similar to the stories shown previously, one of these four stories shown to the audience was ranked by the model as portraying a personality close to the "true" personality of the target character, one was ranked lowly, one was ranked medium, and the final one was a randomly chosen story. Subjects were then asked which story they thought most closely fit the given character's personality.

OCEAN Quality	Question
Openness	Tom finds creative solutions to problems.
Openness	Tom tends to analyze possible outcomes of his plans.
Openness (R)	Tom has difficulty coming up with excellent plans.
Openness	Tom has excellent ideas.
Openness (R)	Tom's ideas are ordinary and hardly unique.
Extraversion	Tom finds it difficult to approach others.
Conscientiousness	Tom gets things done quickly.
Neuroticism	Tom changes his mood a lot.

Table 3
Statements Evaluated for testing Openness

6. Results

In this section we will review the results of our experiments on identifying the protagonist personality traits and choosing stories that align with the protagonist's personality type. For these experiments, we collected results for 176 subjects using Prolific, with each subject randomly assigned to one of the four treatment categories. The category with the smallest number of subjects was Low Openness, which had 35 subjects. The category with the highest number of subjects was High Openness, with 48 subjects.

6.1. Identifying Protagonist Personality Traits

Recall that our first hypothesis was that participants should be able to identify if Tom exhibits either high or low Agreeableness or high or low Openness depending on the treatment group. To do this, we evaluated each user's responses to the statements related to their treatment group. For each statement related to the aspect of personality we were analyzing, we aligned the statements with High-attribute, or Low-attribute implications, ie "Reverse" implications to the statement. For the subjects that fell into a high-categories, we considered it a success if the subjects ranked non-Reverse statements with "Strongly Agree" or "Agree," and Reverse statements with "Disagree" or "Strongly Disagree." Likewise, for the low-categories success was determined if the subjects ranked non-Reverse statements as "Disagree" or "Strongly

	5-Pt Likert Scale		Story Selection	
	p-value	Effect Size	p-value	Effect Size
HO	0.869	0.367	4.412e-08	0.625
LO	0.998	0.303	0.999	0.057
HA	3.941e-13	0.634	2.584e-3	0.447
LA	5.218e-41	0.830	1.752e-09	0.674

Table 4
Experiment Results Individually

	5-Pt Likert Scale		Story Selection	
	p-value	Effect Size	p-value	Effect Size
O	0.988	0.347	4.395e3	0.385
A	1.818e-47	0.731	2.138e-10	0.559

Table 5
Experiment Results Combined

Disagree" and ranked Reverse statements as "Agree" or "Strongly Agree."

To determine if there was a significant effect, we used a binomial exact test, testing the distribution of observed successes and failures against a null hypothesis of users providing random responses to each statement. The results of this analysis are summarized in Tables 4 and 6 under the heading “5-Pt Likert Scale.” Table 4 contains information on each individual treatment, and Table 6 contains results if treatments were aggregated based on either Agreeableness or Openness.

The binomial tests indicate that people are able to correctly identify when the protagonist of the story exhibits high agreeableness (p = 3.941e-13) and low agreeableness (p=5.218e-41). We did not observe overwhelming evidence that participants could identify when Tom exhibited either high or low openness. When taken in aggregate, however, we did find significant differences between how users would respond in both the Openness and Agreeableness categories. These results mostly agree with the results obtained by Shirvani and Ware.

6.2. Choosing Stories According to Personality Type

The second hypothesis we test is the idea that when subjects choose a story that they feel best fits the character’s personality, they will choose the one that our model claims is closest to the original "true" story in personality.

As before, we use a binomial exact test to analyze whether participants select the correct story more frequently than the null hypothesis of random story selection. Binomial tests on our story selection experiments indicate that participants are able to identify stories where Tom exhibits high openness (p=4.412e-08), high agreeableness (p=2.584e-3), and low agreeableness (p=1.752e-09).

7. Discussion

By simplifying and reconstructing the metrics used in the Shirvani19 model, we could provide a personality implementation framework that is easily applied to a wide variety of projects. Having a small-scale framework for personality in narrative planning could be used to enhance other projects, or as a personality framework with which to test personality-adjacent features. Checking that a supposedly multi-personality feature relating to say memory or character beliefs actually works with multiple personality systems requires having access to other systems of personality to use.

	5-Pt Likert Scale		Story Selection	
	p-value	Effect Size	p-value	Effect Size
O	0.072	1.160	0.026	1.60
C	**0.016**	**1.160**	**0.001**	**1.73**
E	0.024	1.167	0.014	1.61
A	**0.048**	**1.167**	**<0.001**	**2.80**
N	0.063	1.128	0.002	2.04

Table 6
Shirvani’s Results

While our work has managed to adapt the Agreeableness metrics to an acceptable degree, we had much less success with Openness. One possible cause is that calculating intellect by calculating the opposition to the character’s plans weighs too closely to Agreeableness. It is also possible that our variance metric for openness punishes plans where the character happens to take the same kind of action regardless of whether the action is the smartest thing to do. It’s also possible that in stories where one character takes few actions compared to other agents, the variance score for openness sees this as showing more variety in the character’s actions simply because the character may not have repeated the same type of action, even if this story shows the character as non-proactive. Alternative approaches to Openness might find more luck in the future, or alternative data-collecting information might enable calculating Shirvani19’s openness metric without issue.

It should also be noted that unlike the original experiment set, in the case of Agreeableness our story selection did much worse than our Likert scale tests, which is the opposite of what Shirvani and Ware found. One explanation could be that we didn’t account for any personality metrics apart from the target score, and thus there were other factors that the audience considered more pertinent than we did. Our attempts to make them consider the story from multiple perspectives may have increased the effect if that is the case.

If we were to test this hypothesis in the future, we would need to expand our generation of alternative stories to check all personality values, not just the targeted ones, and select our "nearest-fit" stories to be ones where the model claims the target character shows a moderate personality in all aspects except for the aspect being tested.

Another possible cause is that the subjects might have selected most likely stories based on the actions of other characters outside of our target character, instead of focusing on Tom’s behavior alone. Although the original four set of stories shown portray the various characters apart from Tom acting in various different ways, some readers might have still attributed their personalities in their selection of most-likely stories. This could be corrected in future studies by simply replacing the names of these characters in the second set of stories, so that the audience views them as different. Alternatively, another story domain with a single character present could be used for future testing.

8. Conclusion

In our attempts to refine the 2019 OCEAN-based personality model into a format that can be applied to story generation tasks as well as story evaluation tasks while still remaining a small-scale easily implemented personality model, we have had some successes and some failures. Our results indicate that of the two OCEAN attribute metrics we sought to refine, only Agreeableness has been properly adjusted into a format that audiences will recognize. Our work on Openness needs to be redefined, and one of the testing metrics we have used should likely be refined as well before using it again.

One problem is that we were focused too intensely on translating the original metrics into runtime-calculable forms, and as such did not reevaluate if alternative solutions might work better. For a first attempt this is still a crucial step to reach, but there are still clear problems. For example, take the original concept of evaluating "intelligence" by way of evaluating the likeliness of other characters opposing the plan as a probability of success. Our adaptation was simply using another measurement of the number of characters likely to oppose the plan, but this results in punishing cases where intelligent characters can act coldly, or manipulative of others.

Still, our work has helped progress towards a personality model that may not be the most refined nor even the most accurate, but could be applied easily and quickly to any given narrative planner for character enhancement or comparative study with other personality models. Testing a baseline model for comparison is a practice seen in countless scientific fields, and providing a model that can serve as one for personality modeling would benefit many future researchers.

References

[1] P. Gervás, B. Lönneker-Rodman, J. C. Meister, F. Peinado, Narrative models : Narratology meets artificial intelligence, 2006. URL: https://api.semanticscholar.org/CorpusID:89613631.

[2] S. Imabuchi, T. Ogata, Story generation system based on propp theory as a mechanism in narrative generation system, in: 2012 IEEE Fourth International Conference On Digital Game And Intelligent Toy Enhanced Learning, 2012, pp. 165–167. doi:10.1109/DIGITEL.2012.47.

[3] F. Peinado, P. Gervás, Creativity issues in plot generation (2005).

[4] A. Shirvani, S. G. Ware, A plan-based personality model for story characters, in: AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, 2019, pp. 188–194.

[5] E. S. de Lima, B. Feijó, A. L. Furtado, Adaptive storytelling based on personality and preference modeling, Entertainment Computing 34 (2020) 100342. URL: https://www.sciencedirect.com/science/article/pii/S187595211930076X. doi:https://doi.org/10.1016/j.entcom.2020.100342.

[6] P. Tambwekar, M. Dhuliawala, L. J. Martin, A. Mehta, B. Harrison, M. O. Riedl, Controllable neural story plot generation via reward shaping, in: Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-2019, International Joint Conferences on Artificial Intelligence Organization, 2019. URL: http://dx.doi.org/10.24963/ijcai.2019/829. doi:10.24963/ijcai.2019/829.

[7] A. Shirvani, S. G. Ware, L. J. Baker, Personality and emotion in strong-story narrative planning, IEEE Transactions on Games 15 (2023) 669–682. doi:10.1109/TG.2022.3227220.

[8] R. Evans, E. Short, Versu—a simulationist storytelling system, IEEE Transactions on Computational Intelligence and AI in Games 6 (2014) 113–130.

[9] e. a. McCoy, Joshua, Social story worlds with comme il faut. (2014) 97–112.

[10] J. C. Bahamón, R. M. Young, A choice-based model of character personality in narrative, in: Workshop on Computational Models of Narrative, 2012, pp. 164–168.

[11] J. C. Bahamón, R. M. Young, An empirical evaluation of a generative method for the expression of personality traits through action choice, in: Proceedings of the Thirteenth AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, AIIDE'17, AAAI Press, 2017.

[12] M. Kreminski, M. Dickinson, M. Mateas, N. Wardrip-Fruin, Why are we like this?: Exploring writing mechanics for an ai-augmented storytelling game, in: International Conference on the Foundations of Digital Games (FDG '20), 2020.

[13] H. Rashkin, A. Bosselut, M. Sap, K. Knight, Y. Choi, Modeling naive psychology of characters in simple commonsense stories, 2018. URL: https://arxiv.org/abs/1805.06533. arXiv:1805.06533.

[14] D. Sander, Models of emotion: the affective neuroscience approach, The Cambridge Handbook of Human Affective Neuroscience (2013) 5–53.

[15] S. G. Ware, C. Siler, Sabre: A narrative planner supporting intention and deep theory of mind, in: AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, volume 17, 2021, pp. 99–106.

[16] C. Deyoung, L. Quilty, J. Peterson, Between facets and domains: 10 aspects of the big five, Journal of personality and social psychology 93 (2007) 880–96. doi:10.1037/0022-3514.93.5.880.