Proceedings, The Thirteenth AAAI Conference on
Artificial Intelligence and Interactive Digital Entertainment (AIIDE-17)

An Empirical Evaluation of a Generative Method for the Expression of Personality Traits through Action Choice

Julio Cesár Bahamón
Department of Computer Science
UNC Charlotte
Charlotte, NC, USA, 28233
jbahamon@uncc.edu

R. Michael Young
School of Computing and
The Entertainment Arts and Engineering Program
University of Utah
Salt Lake City, UT, 84112
young@eae.utah.edu

Abstract

The presence of interesting and compelling characters is an essential component of effective narrative. Well-developed characters have features that enable them to significantly enhance the believability and quality of a story. In this paper, we describe the results of an experiment to evaluate a planning-based narrative generation system that focuses on the generation of stories that express character. The system is designed to automatically produce narratives that show character personality traits through the choices characters make when selecting the means by which they achieve their goals. Results from our study support the hypothesis that an audience presented with stories generated by *Mask* will attribute personality traits to the story characters that have significant correlation with the computational model of personality used to drive the characters' choices.

Introduction

One of the key components of effective narratives is the presence of interesting and compelling characters. Well-developed characters have features that enable them to significantly enhance the believability and quality of a story. In this paper we describe an experiment to evaluate a generative model of plot that facilitates the inclusion of compelling character personality in narratives automatically generated by a planning-based system. In this model, the expression of personality is operationalized as behavior that results from choices made by a character in the course of a story. This operationalization uses the taxonomy defined in the Five-Factor Model (Goldberg 1990) and results from behavioral psychology that link behavior to personality traits (Mehl, Gosling, and Pennebaker 2006). To contextualize the evaluation, we include a high-level description of the system and its core algorithm here as well.

The model focuses on the role that actions performed by characters play over the course of a narrative in the construction of the mental model of the story that the audience forms when experiencing it. Consequently, the generative method is designed to ensure that the story structure and its contents promote the existence of choices for character's actions and make such existence evident to the audience by including contrasting options available to the character. An intelligent mechanism selects the actions that characters perform ensuring their consistency with personality traits assigned to the character by a human author prior to story generation. Results from our study support the effectiveness of this approach for the portrayal of personality traits that are linked to the agreeableness personality factor.

Related Work

A considerable body of research has been dedicated to the development of narrative generation systems and intelligent virtual agents. Among these, several approaches have focused on the portrayal of personality traits. However, these and similar efforts tend to deal primarily with characters as individual agents who are able to react to their immediate environment or to the actions of other agents. Although existing methods have proven successful at specific aspects of narrative generation (Riedl and Young 2010), drama management (Riedl and Stern 2006), portrayal of emotions in virtual agents (André et al. 2000), and natural language generation (Mairesse and Walker 2007), they typically handle very specific and localized elements of narrative without reasoning about the story in its entirety. In contrast, the narrative generation algorithm evaluated in this paper considers characters in the context of an entire story.

The computational model of *narrative* used in this research is based on a modification of the Glaive state-space planning algorithm developed by Ware and Young (2014). Additionally, our work has its foundation on previous research conducted by Young and his colleagues on the use of planning algorithms for narrative generation (e.g., (Harris and Young 2009; Riedl and Young 2010)). The use of AI planners to generate stories was first introduced in systems such as Tale-Spin (Meehan 1977). Considerable effort has been dedicated since then to the development and improvement of techniques, algorithms, and architectures to enable the application of the problem solving capabilities of planners to the automatic generation of narrative that is both interesting and coherent (Riedl and Young 2003). These include the work of Lebowitz on UNIVERSE (Lebowitz 1984), and the IPOCL planning algorithm by Riedl and Young (2010). IPOCL focuses on character intentionality by identifying goals that explain a character's actions, which is done without considering character personality. In contrast, our work focuses on the selection by a character of specific

actions to achieve her goals as the means for the system to portray distinct personality traits. We envision that both approaches can be complementary.

Other approaches have focused on systems that direct the interaction among story characters. Work by Assanie (Assanie 2002) on the extension of synthetic characters based on the Soar QuakeBot environment (Laird and Jones 1998) provided agents with the ability to adjust to changing goals provided by an external narrative manager. Work by Riedl and Stern on drama managers focused on semi-autonomous agents able to avoid situations in conflict with their goals and behave in a way that justifies failure when a conflict occurs (Riedl and Stern 2006).

Our computational model of *character personality* is based on the Five-Factor Model (FFM), or Big Five (Goldberg 1990). The FFM provides a taxonomy for the classification of personality using the factors Openness, Conscientiousness, Extroversion, Agreeableness, and Neuroticism. Each factor contains distinct bi-polar personality traits, e.g. generosity vs. selfishness. Traits can in turn be mapped to specific behavior (Funder and Sneed 1993; Mehl, Gosling, and Pennebaker 2006). We focus on Agreeableness because this factor is commonly used to describe *good* vs. *evil* personalities (McCrae and John 1992), a component of character types and themes frequently used in narrative. A number of research efforts have been aimed at the operationalization of the FFM to create richer and more expressive characters. The work of Doce *et al.* (2010) applied the FFM to create distinguishable personalities in virtual agents by using personality traits to affect specific cognitive and behavioral processes (e.g. coping mechanisms, bodily expressions). Work by Bouchet and Sansonnet (2011) applied the FFM and a model of cognition to implement a system in which rational agents make decisions after reasoning about goals modulated by pre-assigned personality traits. Other work has focused on a specific subset of character actions: utterances in dialog. Of particular interest to our research is PERSONAGE (Mairesse and Walker 2007), a natural language generator that produces dialog to meet predefined personality requirements. PERSONAGE uses the FFM to create a mapping between personality traits and dialog utterances.

The *Mask* System

The *Mask* narrative generation system has its foundation on a character-centric model of action selection. Within this model, we assume that story characters act intentionally and consider multiple courses of action as they pursue their goals. Consequently, we use an operationalization of character intentionality based on the one defined by Riedl and Young for the IPOCL algorithm (2010) and later expanded by Ware and Young for the CPOCL algorithm (2014). In these approaches, sub-plans called *intention frames* explicitly mark the adoption of a new goal by a character, the action in the story that gives rise to the new goal, the steps that the character takes in pursuit of that goal and the point in the story at which the sub-plan to achieve the goal ends execution.

In contrast to a traditional narrative planner where a uniform planning process handles the construction of a story, in choice-based narrative generation, planning is interleaved between a macro-planner and a micro-planner (MP). The macro-planner is responsible for constructing the authorial view of the global story and its causal coherence. The micro-planner is a restricted version of a planning system, designed to simulate the character's decision-making process as she works toward accomplishing a goal. To represent the character's process of considering multiple plausible courses of action to achieve her goal, the MP takes a snapshot of the current state and generates the set of possible intention frames that allow the character to achieve such goal and then evaluates them utilizing a simplified model of affect appraisal theory. This facilitates the implementation of a process where character choice is treated as a first class object, and as such it guides the construction of the story.

The determination of a character's choice for action is triggered when a character forms an intention to achieve a new goal. When this occurs, the MP generates, evaluates, and ranks the set of intention frames that can be constructed to achieve the character's goal. Note that in our model not all new intentions will result in a choice. For example, even though multiple intention frames for achieving a character's new goal may exist, their effects on the affective response of other story characters may not have measurable differences relative to one another. When this condition occurs, one of the intention frames is chosen non-deterministically.

We say that a *choice* occurs just when a character adopts a new intention whose achievement requires the character to chose between at least two contrasting intention frames. *Mask* chooses between the alternate intention frames using a simplified version of the appraisal theory proposed by Lazarus (1991). Our model represents the impact of a character's choice for action on the emotions of other characters in terms of the beneficial or harmful appraisal of its effects, i.e., how it changes the state of the story world. This evaluation centers on whether a character's course of action supports or prevents other characters from accomplishing their own goals. The MP evaluates every intention frame using the following criteria: (1) its effects on the current intentions of other characters, (2) its effects on the current intentions of the character who is making a choice, and (3) the personality of the character who is making the choice. For example, a non-agreeable character will be more likely to make choices that prevent other characters from achieving their goals. The same mechanism can be used to model conscientiousness; however, the model would need to be extended to model other FFM factors.

When a choice is identified, i.e., two or more contrasting courses of action to achieve the same goal exist, two distinct intention frames are selected: the *choice*, which is the intention frame that the character will execute, and the *contrast*, which is the intention frame used to highlight what the character could have done but chose not to do. Steps in the *choice* intention frame are returned to the macro-planner to specify the course of action it then constructs for the given character, merging these actions into the global story structure it generates. Steps in the *contrast* intention frame are not added to the plan, but are recorded to support discourse devices that communicate the character's deliberation process.

Algorithm Overview

The *CB-Glaive* narrative generation algorithm used by *Mask* is based on a modification of the Glaive algorithm (Ware and Young 2014), using a technique based on previous work by Bahamón, Barot and Young (2013; 2015), to include a mechanism that evaluates the effect of a character's actions on the individual intentions of other story characters. This represents the character's process of considering multiple plausible courses of action to achieve a goal. Instead of using a non-deterministic method to select action schemata during plan construction, the mechanism utilizes individual character intentions and the personality traits of the protagonist to guide the selection. The mechanism also treats the goals defined in the planning problem differently from the individual goals of story characters.

CB-Glaive makes use of a state-space plan-based story representation, where STRIPS-like action schemata (Fikes and Nilsson 1971) describe the set of actions available in the story world. An **Action Schema** is a template for an action possible in the story world, described by a tuple $\langle \mathsf{ActionType}, \mathsf{Pre}, \mathsf{Eff}, V, \mathsf{MainChar} \rangle$ where ActionType is a unique identifier for the action, Pre is a set of literals that must be true prior to executing the action (preconditions), Eff is a set of literals established by the execution of the action (effects), V is the list of variables used in the template, and MainChar is one of the variables in V that designates the story character who performs the action.

CB-Glaive first identifies when a character forms a new intention, i.e., adopts a new goal to pursue, and determines if there exist multiple ways for such goal to be achieved. When this is the case, a choice point exists and the alternate versions of the story produced by each viable course of action are explored and analyzed. The result is a ranking of each alternate version based on how consistently the effects of steps included in each story version portray the personality traits of the character who is making the choice. The story version with the most consistent effects is selected and the action schemata it contains are selected to be added to the plan after the choice point. Additionally, the mechanism makes sure that a story version whose effects are least consistent with the personality of the character is a possible path that the story could follow as well. Finally, the planner records information about the alternate but not followed path for later use by a discourse manager; this is designed to help ensure that the information necessary to convey the existence of a choice can be presented to the audience.

CB-Glaive maintains a database of current intention frames. This information guides plan construction and informs the reasoning mechanism utilized by the micro-planner to consider choice alternatives. The information recorded about every choice includes: (1) the step in the macro-plan where the choice was identified, (2) the *character name* that represents the story character who made the choice, (3) the intention frame that was selected by the micro-planner as most consistent with the character's personality, and (4) the intention frame that was selected by the micro-planner as least-consistent with the behavior exhibited by the character in (3). Note that this additional bookkeeping is only done when a choice is detected. The pseudo-code for CB-Glaive is provided in Algorithm 1.

The initial invocation of CB-Glaive is similar to that of Glaive (Ware and Young 2014). The values in $\langle \Pi, P, \sigma, G, X, \mathsf{Choices} \rangle$ are initialized to represent the initial and goal states for the planning problem, the set of story characters, the initial set of character goals (which may be empty), and the set of unexplained steps, which is initially empty. Note that character goals may be assigned by the author or adopted as the story is constructed. Additionally, Choices is initialized to $\{\{s_0, \emptyset\}, \{s_G, \emptyset\}\}$, to indicate the lack of choices before the initial plan step and after the final plan step. On subsequent invocations of the algorithm, if all of the author's goals have been satisfied and there are no unexplained steps, the process succeeds and a plan is returned. Otherwise, CB-Glaive selects a potentially motivated step (s) whose preconditions are satisfied in the current state (σ) and adds it to the plan.

The selection method uses a heuristic that favors steps which are present in a character's chosen course of action, i.e., they are part of a *choice* intention frame previously evaluated and selected by the MP as most consistent with a character's personality factor. If the newly added step adds a new character goal in one of its effects (it establishes a new intention) and the main character who adopts the goal has been assigned a personality factor, the MP will be invoked. The MP will evaluate the different courses of action that the character can choose to accomplish the goal and whether these exhibit any contrast with respect to her personality. If at least two contrasting frames are found, these are stored in Choices for later use by the heuristic and the discourse generator. From this point on, CB-Glaive operates in the same manner as Glaive, which mainly includes keeping track of steps that require explanation in the form of character intentions and the bookkeeping needed to keep track of such intentions.

Evaluating *Mask*

We provide an empirical evaluation of the *Mask* narrative generation system. In this evaluation, stories were automatically generated by the planning-based algorithm and translated into natural language using the *SimpleNLG* natural language realizer (Gatt and Reiter 2009). These stories were then presented to a human audience, who were asked to read the story and then answer a series of questions about the personality of the story's main character.

Data collected from the study aimed at testing the validity of the computational model used in the implementation of *Mask* to generate stories in which specific actions and action sequences are selected to portray character personality traits. The model, described briefly above, is designed to promote character actions that portray the characters as making choices for action and action sequences that are consistent with a pre-assigned personality factor. In this study, we focused on the Agreeableness factor as defined by Goldberg in the Big-Five Factor Model (Goldberg 1990). Even though it could be argued that *Mask* provides a framework for the portrayal of a diverse set of personality traits, the results that will be discussed below focus on the successful portrayal of behavior that is associated with agreeableness. The imple-

Algorithm 1 *CB-Glaive(*$\Pi, P, \sigma, G, X,$ Choices*)*

Where Π is a totally ordered, fully-ground plan, P is a planning problem $\langle \Lambda, S_0, S_G, C \rangle$, σ is the current state, G is the set of character goals, X is the set of unexplained steps (not in an intention frame), and Choices is the set of choices made by characters in the course of the story.
Nondeterministically choose a potentially motivated step s whose preconditions are satisfied in σ. Assign a higher priority to steps in Choices.
▷ **Update the plan**
Add step s to Π.
Apply the effects of s to σ.
for each effect of s like c *intends* g **do**
    Add a new character goal $\langle c, g \rangle$ to G.
    ▷ **Check if a choice exists**
    if c has been assigned a personality factor by author **then**
        Invoke MP, evaluate c's choices for action (affect appr.)
        Add intention frame returned to Choices
        Store contrasting intention frame
    end if
end for
if any characters consent to s **then** Add s to X
end if
for each character goal $g = \langle c, g \rangle \in G$ **do**
    for each intention frame I for c ending in g **do**
        Remove g from G.
        for each step $t \in p$ **do**
            if t is explained **then**
                Remove t from X for all nodes*
            end if
        end for
    end for
end for
if any node* satisfies the author's goals and $X = \emptyset$ **then**
    return Π for that node ▷ **Solution found**
else
    CB-Glaive($\Pi, P, \sigma, G, X,$ Choices*)*
end if

mentation and evaluation of behavior typical of other other factors will be the subject of future work.

The main objective of the study was to evaluate the following thesis: **An audience presented with a set of alternate action sequences generated for a story character by the *Mask* planning-based narrative generation system will select personality trait ratings that have significant correlation with the choices made by such character**. Additionally, the experiment was also designed to test whether an audience can recognize when a character's behavior over the course of a story is inconsistent with her assigned personality factor. This evaluation was conducted using stories generated automatically by *Mask*, but hand-selected from the range of solutions to our domain's planning problem because *Mask*'s computational model characterized them as having action sequences representative of the personality factor assigned to the protagonist, in this case *agreeableness*.

Experimental Design

In order to determine whether readers are able to make inferences about the personality traits of characters by observing them make choices corresponding to those that *Mask* is designed to generate, we conducted a study where participants read a series of short stories produced by *Mask* and selected because they include events where the main character is faced with having to choose a specific course of action from a set of applicable alternatives. These stories are presented in natural language (English), that results from running the plan-based story structures produced by *Mask* through a Natural Language realizer. The chosen action and its alternatives contrast with each other, ensuring the portrayal of behavior that was either typical of the agreeableness personality factor or its opposite. The mechanism used to generate the natural language representation of the stories is one of the components implemented as part of the *Mask* system. This mechanism uses a simple discourse manager that is built on top of the *SimpleNLG* Java API (Gatt and Reiter 2009).

Subjects were assigned in a round-robin manner into one of six groups. Each participant read three stories, but the ordering of the stories was counterbalanced across the six groups to control for potential learning effects. Three of the groups, called the experimental condition groups, read three stories characterized as follows: a) a story that includes actions where the main character makes choices to act in a manner that is consistent with an agreeable personality, b) a story that includes actions where the main character makes choices to act in a manner that is not consistent with an agreeable personality, and c) a story that includes actions where the main character makes choices to act in in a manner that is not consistent with an agreeable personality, except for the last action in the story. The last action is designed to be inconsistent with the character's non-agreeable personality, i.e., it demonstrates traits from an agreeable personality (e.g. altruism). The stories used are named and characterized in Table 1.

ID	Story Name	Personality Factor	Sentences	Chars
S1-1	Escape from The Forest Version 1	Treatment High Agreeableness	20	3
S1-2	Escape from The Forest Version 2	Treatment Low Agreeableness	20	3
S1-3	Escape from The Forest Version 3	Control Neutral	22	3
S2-1	The Kidnapping of The Princess Version 1	Treatment High Agreeableness	20	4
S2-2	The Kidnapping of The Princess Version 2	Treatment Low Agreeableness	20	4
S2-3	The Kidnapping of The Princess Version 3	Control Neutral	20	4
S3-1	Jane Version 1	Treatment Inconsistent Choice	19	2
S3-2	Jane Version 2	Control Consistent Choice	19	2

Table 1: Story Versions

Participants in the other three groups, called the control condition groups, were presented with two stories where the choices for action of the character are designed to be neutral and a version of the story where the character's last action is consistent with the character's non-agreeable personality, i.e., it demonstrates traits from a non-agreeable personality. Participants' assignments to one of the six possible groups are shown in Tables 2 and 3. The order of the columns in each table indicates the order in which stories were pre-

sented to the participant.

Random Assignment of Participants	First Story			Second Story			Third Story	
	S1-1	S1-2	S1-3	S2-1	S2-2	S2-3	S3-1	S3-2
Experimental Group 1-A	X				X		X	
Experimental Group 2-A		X		X			X	
Control Group A			X			X		X

Table 2: Participants Who Read Story #1 First

The stories used were based on passages from *The Hobbit: Chapter 5 - Riddles in The Dark* (Tolkien 2001), *The Life of Lazarillo of Tormes; his fortunes and misfortunes as told by himself, Chapter 1* (Rudder and Puertolas 1973), and also on themes typically found in popular fairy tales. The names of characters, places, and other story content were modified to obscure the actual narrative sources in order to prevent the introduction of bias due to prior knowledge of the story content. These stories had a common structure, typical of fairy tales and popular short stories: (1) a brief introduction leading to a situation involving the main character, (2) followed by a moment where the main character must make a decision to resolve the situation, and (3) the story end that results from the character's decision. Most of the story content was identical in all versions of each story. Two or three sentences toward the end of each story were different, depending on the personality factor that was being portrayed. The stories all involved characters with personality traits either typical (e.g. altruism, unselfishness) or atypical (e.g. belligerent, selfish) of the agreeableness personality factor. Each of the stories required the main character to choose between one of several courses of action. There were three versions of each story, depending on the choice made by the character. In the treatment versions, the choice portrays behavior that is either typical of high agreeableness or typical of low agreeableness. The control versions eliminated the need for the character to make a choice through the intervention of a neutral third party or event.

The study included a total of 96 participants, recruited from NC State University and through social media. Only participants who completed the experiment in its entirety are included in the results we describe in this paper. Subjects accessed the system via a web browser and were presented with textual narratives followed by a series of questions with Likert-scale responses. Participants completed an initial demographic information survey, then read each story, completing a personality survey about the story's main character before reading the next story. The personality survey collected information used to measure the participants' perception of the main character's personality traits in the story they had just read. Subjects provided Likert scale responses

Random Assignment of Participants	Second Story			First Story			Third Story	
	S2-1	S2-2	S2-3	S1-1	S1-2	S1-3	S3-1	S3-2
Experimental Group 1-B	X				X		X	
Experimental Group 2-B		X		X			X	
Control Group B			X			X		X

Table 3: Participants Who Read Story #2 First

Agreeableness Score - First Story					
	DF	Sum Sq	Mean Sq	F value	PR(>F)
Story-ID	2	10.99	5.494	12.71	1.32e-05
Residuals	93	40.21	0.432		
Agreeableness Score - Second Story					
	DF	Sum Sq	Mean Sq	F value	PR(>F)
Story-ID	2	6.76	3.379	8.946	0.00028
Residuals	93	35.13	0.378		
Agreeableness Score - Combined Dataset (S1 and S2)					
	DF	Sum Sq	Mean Sq	F value	PR(>F)
Factor Portrayed	2	15.03	7.516	12.71	6.62e-06
Residuals	189	111.75	0.591		

Table 4: ANOVA Test Results for the Agreeableness Factor

to questions about the stories' main character that were drawn from the Big Five Inventory (BFI) instrument (John, Donahue, and Kentle 1991; Benet-Martínez and John 1998; John, Naumann, and Soto 2008) developed at the Berkeley Personality Lab. Post-processing of the data was done following the instructions provided by the instrument's authors. The post-processed results of the survey provided a score of the story protagonist's personality factors (e.g agreeableness). Even though the raw experiment data was discrete, the score data was a continuous value between 1 and 5.

Results and Statistical Analysis

We utilized an Analysis of Variance (ANOVA) statistical test to analyze the results of the experiment. The test was applied on the data obtained from the agreeableness score with respect to the story version that participants read. Results from the test are shown in Table 4. The test indicates statistical significance such that the null hypothesis can be rejected in the case of the agreeableness personality factor. This provides evidence that there is correlation between the version of the story that participants read and the perceived personality factor of the main character in such story.

Measuring the Consistency of Character Behavior

The third story that was presented to the study participants was designed to evaluate whether they could perceive the existence of inconsistent character behavior. In this story, the main character performs several actions that are typical of a low-agreeableness personality factor (e.g. stealing). Toward the end of the control version of the story (S3-2), the character's final action is consistent with her personality factor, i.e., she behaves according to the personality traits typical of low-agreeableness. In the treatment version of the story (S3-1), the character's final action is not consistent with her personality factor. The action she performs is an example of altruism, which is a trait typical of a high-agreeableness personality factor and hence in contrast with the personality portrayed by the character up to that point in the story.

For Story S3-1, **41.54%** (27 out of 65) reported perceiving inconsistency in the main character's behavior, while **58.46%** of subjects (38 out of 65) reported not perceiving inconsistency. For Story S3-2, **70.97%** (22 out of 31) reported perceiving inconsistency in the main character's behavior, while **29.03%** (9 out of 31) reported not perceiving inconsistency. Although these results are not quite as strong

as the ones obtained in an earlier formative study (Bahamón and Young 2016), the percentage of responses from participants within each group who reported inconsistent behavior indicates that such behavior was somewhat perceivable by the audience; however, this is an informal assessment and there is no conclusive evidence to make any claims. A larger percentage of the participants who read the control version of the third story (S3-1) indicated that the main character did not behave in a consistent manner, whereas the opposite was true for participants who read the treatment version (S3-2). To further evaluate this aspect, participants were specifically asked if they thought that the character's actions regarding a specific character action seemed odd or unusual, with similar results.

Even though the results from this portion of the experiment are not statistically significant, they lend some support to our thesis, by strengthening the argument that audiences are able to detect contrast between different character actions when such contrast exists. Nonetheless, it is confounding that the contrast perception is closer to the opposite of what the story was designed to portray. It is worth noting that informal feedback provided by some of the study participants indicated a possible ambiguity in the survey question that was used to collect this data. It is possible that the story accomplished its objective but that the wording of the question was not clear enough to obtain accurate results.

We also analyzed the consistency data using the *Pearson's Chi-square test of independence with Yates' continuity correction*. The results yielded a χ^2 value of **6.5124**, with 1 degree of freedom and a p-value of **0.01071**. This allows us to reject the null hypothesis that the perceived consistency of the story is independent from the story version read. Furthermore, these results provide support for our hypothesis that given a sequence of actions, performed by the same character, an audience will be able to perceive that a character's actions are inconsistent with the personality factor that is exhibited in the character's previous actions. However, it is very important to note that the audience did not perceive the consistency of the stories in the manner that we expected they would, i.e., inconsistency in the story character's choice for action was perceived, but this occurred on the story in which the character's actions should have been perceived as consistent. Conversely, consistency was perceived in the story in which the character's actions should have been perceived as inconsistent. A likely source for the ambiguity of these results is in the possible lack of clarity in the survey questions used. It is also plausible that the content of the stories used may have proved confusing to the audience. We believe that this question deserves further exploration.

Personality Measurement The study participants also completed a personality evaluation of the main character in the third story (S3). This evaluation was conducted using the BFI instrument described earlier. Because both versions of the story were designed to portray the main character as having a low-agreeableness personality factor, we expected scores to be in the 1.0 to 3.0 range. The average personality score for the Agreeableness factor of this character was **3.26** for the treatment version of the story (S3-1) and **1.61** for the control version (S3-2). The score for S3-2 is clearly indicative of a low-agreeableness personality factor. This is precisely what the story produced by *Mask* was designed to portray. Also, note that the score is significantly higher (103%) for the story version where the character performed exactly one action that was consistent with a high-agreeableness factor. This particular result is something we expected, yet it is surprising considering the results from the consistency perception data because it indicates that actions of the main story character elicited the perception of different personality traits depending on the version of the story that was read. Nonetheless, the audience did not appear to identify that there was an inconsistency in the character's actions, even though it is this very inconsistency that causes the significant difference in the personality scores.

We also utilized an ANOVA test to analyze the personality data collected for Story #3. The test was applied on the data obtained from the agreeableness score with respect to the story version that participants read. Results from the test are shown in Table 5.

	DF	Sum Sq	Mean Sq	F value	PR(>F)
Story-ID	1	57.67	57.67	108.2	<2e-16
Residuals	94	50.08	0.53		

Table 5: ANOVA Results for Agreeableness - Third Story

Discussion

Analysis of the data collected from the evaluation of the *Mask* system indicates that the personality scores for the **agreeableness** factor and their relationship with the story read by the participant are statistically significant in favor of rejecting the null hypothesis. The results support our thesis that an audience presented with stories generated by *Mask* will attribute personality traits to the story characters that have significant correlation with the computational model of personality used to drive the characters' choices. Additionally, an analysis of the agreeableness scores for each story, independently from each other, yielded results that disprove the existence of an imbalance between the results for each story. Each story had the expected effect on the audience when considered on its own and when considered within the context of both stories read. Furthermore, the data collected from this experiment yielded results that are comparable to those obtained from a formative study conducted earlier. This is important since it provides validation that the algorithm can be effectively implemented as part of a planning-based narrative generation system.

Conclusions and Future Work

The results from the experiment provide empirical evidence of the validity of our approach. Validation of the choice-based mechanism used by the CB-Glaive algorithm has been completed also using a fully-automated system in which an AI planner generates the story structure and a discourse manager creates the human-readable text. The results from the evaluation are statistically significant in a manner that supports the hypothesis previously stated that *an audience presented with a set of alternate action sequences for a story*

character will select personality trait ratings that have significant correlation with the choices made by such character. This hypothesis provided the foundation for the design of the *Mask* narrative generation system.

The data obtained from the experiment clearly indicates that the narrative generation approach used in the design of the CB-Glaive algorithm is valid for the portrayal of agreeableness. This is a significant result for the advancement of planning-based narrative generation because it provides a solid foundation for the portrayal of a theme that is used in many narratives, namely the juxtaposition of good (high-agreeableness) vs. evil (low-agreeableness) characters (McCrae and John 1992). Our future work will seek to extend the *Mask* system so that it can produce stories with these properties in ways that are more efficient than typical non-declarative approaches and also to include the portrayal of additional personality factors.

References

André, E.; Klesen, M.; Gebhard, P.; Allen, S.; and Rist, T. 2000. Integrating models of personality and emotions into lifelike characters. *Affective interactions* 150–165.

Assanie, M. 2002. Directable synthetic characters. In *Proceedings of the AAAI Spring Symposium on Artificial Intelligence and Interactive Entertainment*, 1–7.

Bahamón, J. C., and Young, R. M. 2013. CB-POCL: A Choice-Based Algorithm for Character Personality in Planning-based Narrative Generation. *CMN 2013* 4–23.

Bahamón, J. C., and Young, R. M. 2016. A formative study evaluating the perception of personality traits for planning-based narrative generation. In *ICIDS 2016*, 123–135. Springer.

Bahamón, J. C.; Barot, C.; and Young, R. M. 2015. A goal-based model of personality for planning-based narrative generation. In *AAAI*, volume 6, 4142–4143.

Benet-Martínez, V., and John, O. P. 1998. Los Cinco Grandes across cultures and ethnic groups: Multitrait-multimethod analyses of the Big Five in Spanish and English. *Journal of Personality and Social Psychology* 75(3):729–750.

Bouchet, F., and Sansonnet, J.-P. 2011. Influence of personality traits on the rational process of cognitive agents. In *Proceedings of WI - IAT*, 81–88. IEEE CS.

Doce, T.; Dias, J.; Prada, R.; and Paiva, A. 2010. Creating individual agents through personality traits. In *Intelligent Virtual Agents*, 257–264. Springer.

Fikes, R. E., and Nilsson, N. J. 1971. Strips: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence* 2(3-4):189–208.

Funder, D. C., and Sneed, C. D. 1993. Behavioral manifestations of personality: An ecological approach to judgmental accuracy. *Journal of Personality and Social Psychology* 64(3):479–490.

Gatt, A., and Reiter, E. 2009. SimpleNLG: A realisation engine for practical applications. In *ENLG Workshop*, 90–93.

Goldberg, L. R. 1990. An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology* 59(6):1216.

Harris, J., and Young, R. M. 2009. Proactive mediation in plan-based narrative environments. *IEEE TCIAIG* 1(3):233–244.

John, O. P.; Donahue, E. M.; and Kentle, R. L. 1991. The big five inventory - versions 4a and 54. *Berkeley: University of California, Berkeley, Institute of Personality and Social Research*.

John, O. P.; Naumann, L. P.; and Soto, C. J. 2008. Paradigm shift to the integrative big five trait taxonomy. *Handbook of personality: Theory and research* 3:114–158.

Laird, J. E., and Jones, R. M. 1998. Building advanced autonomous ai systems for large scale real time simulations. In *Computer Games Development Conference*.

Lazarus, R. S. 1991. Progress on a cognitive-motivational-relational theory of emotion. *American Psychologist* 46(8):819–834.

Lebowitz, M. 1984. Creating characters in a story-telling universe. *Poetics* 13(3):171–194.

Mairesse, F., and Walker, M. 2007. PERSONAGE: Personality generation for dialogue. In *ACL*, 496–503.

McCrae, R. R., and John, O. P. 1992. An introduction to the five-factor model and its applications. *Journal of Personality* 60(2):175 – 215.

Meehan, J. R. 1977. Tale-spin, an interactive program that writes stories. In *IJCAI*, 91–98.

Mehl, M. R.; Gosling, S. D.; and Pennebaker, J. W. 2006. Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology* 90(5):862–877.

Riedl, M. O., and Stern, A. 2006. Failing believably: Toward drama management with autonomous actors in interactive narratives. *TIDSE* 195–206.

Riedl, M. O., and Young, R. M. 2003. Character-focused narrative generation for execution in virtual worlds. *Virtual Storytelling. Using Virtual Reality Technologies for Storytelling* 47–56.

Riedl, M. O., and Young, R. M. 2010. Narrative Planning: Balancing Plot and Character. *JAIR* 217–268.

Rudder, R. S., and Puertolas, C. C. d. R., eds. 1973. *The Life of Lazarillo of Tormes: His fortunes and misfortunes as told by himself.* New York: Ungar.

Tolkien, J. R. R. 2001. *The Hobbit*. Boston: Houghton Mifflin.

Ware, S. G., and Young, R. M. 2014. Glaive: A state-space narrative planner supporting intentionality and conflict. In *Proceedings of AIIDE-14*, 80 – 86.

Ware, S.; Young, R.; Harrison, B.; and Roberts, D. 2014. A computational model of plan-based narrative conflict at the fabula level. *IEEE TCIAIG* 6(3):271–288.