Initial Results for Measuring Four Dimensions of Narrative Conflict

Stephen G. Ware, Brent Harrison, R. Michael Young, and David L. Roberts
Digital Games Research Center
Department of Computer Science
North Carolina State University
Raleigh, NC 27695

Abstract

Conflict is an essential element of interesting stories. In previous work, we proposed a formal model of narrative conflict. We also described 7 dimensions which can be used to distinguish one conflict from another: *participants, subject, duration, balance, directness, intensity,* and *resolution*. This paper presents the results of an experiment designed to measure how well our metrics for *balance, directness, intensity,* and *resolution* predict the responses of human readers when asked to measure these same values in a set of four stories. We conclude that our metrics are able to rank stories similarly to human readers.

Introduction

Narratologists, screen writers, and other researchers in computer narrative agree that conflict is an essential element of stories (Vale 1973; Ryan 1991; Szilas 2003; Herman, Jahn, and Ryan 2005; Barber and Kudenko 2007; Abbott 2008). It provides an impetus for the action (Egri 1988), structures the discourse (Abbott 2008), and keeps the audience engaged in the unfolding narrative (Gerrig 1993).

Our previous work (Ware and Young 2010; 2011a; 2011b) defined a formal model of conflict based on AI planning. In short, conflict occurs when a goal seeking agent's plan is thwarted by another agent, the environment, or its own plans to achieve other goals. This definition, inspired by research in narratology, is intentionally broad to cover all kinds of conflict.

In order to provide greater control over story content, we identified seven dimensions from various narratological sources that can be used to distinguish one conflict from another. The first three—*participants, subject,* and *duration*—have discrete values which can be directly observed in a story. The other four—*balance*, *directness*, *intensity*, and *resolution*—are continuous values and more subjective. No consensus exists on how to measure these dimensions.

We provided four simple formulas to measure each of these last four dimensions and designed an experiment to test whether the observations of human readers correspond to the values predicted by our formulas. This paper presents the findings of that experiment along with an analysis of the results. We conclude that our formulas for *balance*, *directness*, and *resolution* rank stories in the same order as human readers, and that our formula for *intensity*, while less accurate, still ranks stories similarly to human readers.

This work is an attempt to operationalize a few of the innate story metrics used by human readers into formulas which can be used by machines to evaluate the content of narratives. By capturing a model of how humans evaluate stories, we can guide story generation systems to produce content that is better suited to meet the expectations of the audience.

Related Work

Much previous work exists on modeling human perception with quantitative metrics. Yannakakis (2008) provides a survey of research that measures concepts like *fun* and *flow* in the context of video games. Less work has been done specifically in narrative. Peinado and Gervás (2006) collected four metrics from human readers evaluating the quality of stories produced by their ProtoPropp system: *linguistic quality*, *coherence*, *interest*, and *originality*.

Our approach differs somewhat from these because we wish to measure properties of stories apart from their effects on the reader. The dimensions of conflict answer the *who? what? when?* and *how?* questions; they are designed so that readers can agree on their values even when they do not agree on how fun or interesting a given conflict is.

At least two previous story generation systems have attempted to reason about conflict quantitatively. The IDtension system (Szilas 2003) assigns a "conflict value" to each action in a story that represents the degree to which a character is forced to act against its moral principles. The Mexica system (Pérez y Pérez and Sharples 2001) measures the amount of tension that the reader perceives in the story at each world state, allowing the system to craft a pattern of rising and falling action.

Because conflict is such a diverse phenomenon, we have chosen to measure many individual dimensions rather than attempt to quantify conflict as a single value. This higher level of detail will allow story generating systems to produce content with more specific constraints.

Dimensions of Conflict

Complete formal descriptions for each dimension are given by Ware (2011b). Some essential notation is reproduced here.

We assume that some conflict c exists between character a_1, who intends to carry out a sequence of actions T_1, and character a_2, who intends to carry out a sequence of actions T_2. Some action in T_1 conflicts with an action in T_2—that is, some action in T_1 prevents a_2 from executing one of the actions in T_2. Let E be the set of actions which actually occur in the story. E may contain some actions from both T_1 and T_2, but cannot contain all the actions from both.

Dimensions are measured from some character's point of view. In general, a dimension is expressed as $name(c, a)$ where $name$ is the name of the dimension, c is the conflict, and a is the character from whose point of view the dimension is being measured (either a_1 or a_2).

We also rely on two additional functions with the range $[0, 1]$:

- $\pi(T)$ measures how likely some sequence of actions T is to succeed.
- $\text{utility}(a, T)$ measures how satisfied actor a is with the state of the world after the sequence of actions T occurs. $\text{utility}(a, \emptyset)$ is the character's utility before the conflict begins.

Brief examples from the film series *Star Wars* are provided to illustrate each dimension.

Balance

Balance measures the relative likelihood of each side in the conflict to succeed (regardless of the actual outcome):

$$\text{balance}(c, a_1) = \frac{\pi(T_1)}{\pi(T_1) + \pi(T_2)}$$

The range of *balance* is $[0, 1]$. If a_1 is likely to prevail—that is, $\pi(T_1)$ is close to 1, then balance is high for a_1. If the opposing participant, a_2, is is more likely to prevail, then balance is low for a_1.

When Obi Wan Kenobi challenges Darth Vader to a duel in *Star Wars: A New Hope*, he knows that he cannot win. Vader's skill is at its peak while Kenobi's skill is waning with age. In this conflict, the balance for Kenobi is low while the balance for Vader is high.

Directness

Directness measures how close the participants are to one another:

$$\text{directness}(c, a_1) = \frac{\sum_{i=1}^{n} \text{closeness}_i(a_1, a_2)}{n}$$

3 types of *closeness* are measured in this domain: physical closeness, emotional closeness, and interpersonal closeness. The range of *directness* and each form of *closeness* is $[0, 1]$.

During the climax of *Star Wars: Return of the Jedi*, Luke Skywalker and Darth Vader are physically close because they are engaged in a duel and emotionally close because of their family relation.

Interpersonal closeness is non-zero when one agent participates in the conflict via other agents. Luke is in conflict with the Emperor even before they meet. The Emperor operates through his underlings, putting interpersonal distance between him and Luke.

Intensity

Intensity is the difference between how high a participant's utility will be if she prevails and how low it will be if her opponent prevails:

$$\begin{array}{lcc} best & = & \max\left(\text{utility}(a_1, \emptyset), \text{utility}(a_1, T_1)\right) \\ worst & = & \min\left(\text{utility}(a_1, \emptyset), \text{utility}(a_1, T_2)\right) \\ \text{intensity}(c, a_1) & = & best - worst \end{array}$$

The range of *intensity* is $[0, 1]$. Two factors influence this formula: how much can be gained and how much can be lost. Situations which are high risk or high reward have medium intensity, while situations which are both high risk and high reward have high intensity. Like balance, intensity is measured regardless of the actual outcome of the story.

The Rebel Alliance's plan to destroy the Death Star in *A New Hope* is very intense. If they succeed they will cripple the Empire, but if they fail their rebellion will be crushed.

Resolution

Resolution measures the change in utility a participant experiences after a conflict ends:

$$\text{resolution}(c, a_1) = \text{utility}(a_1, E) - \text{utility}(a_1, \emptyset)$$

The range of resolution is $[-1, 1]$.

Luke and the Rebel Alliance overcome the Empire at the end of *Return of the Jedi*. Their resolution is high, while the resolution for Darth Vader and the Emperor is low.

Design of the Experiment

We designed an experiment to test whether or not the formulas we defined for *balance*, *directness*, *intensity*, and *resolution* can rank stories in the same order as human readers.

Each of the four dimensions being tested can be expressed as a real number between $[0, 1]$ or $[-1, 1]$ for a given conflict in a given story. For example, the dimension of *directness* has the range $[0, 1]$. An indirect conflict might have a directness value of 0.2, while a very direct conflict might have a value of 0.9.

The task of predicting the exact value a reader will report is difficult considering how sensitive these concepts are to subtleties of interpretation. Simply predicting high or low is much easier, but success in this task would provide less support for the strength of our model. We attempt to reach a middle ground by deriving formulas which can order a set of four stories in the same order given by a human reader.

Each participant in the experiment was shown the same four stories (given in figure 1) and asked to rank them from lowest to highest for each dimension. If readers agree on an ordering, and if that ordering agrees with our predictions, we assume that our formulas can approximate these dimensions of conflict.

Table 1: The four dimensions investigated in the study, their formulas, and their descriptions as given to the participants. Descriptions are intended to be short and suitable for a high school reading level.

Dimension	Formula	Description
balance	$\frac{\pi(T_1)}{\pi(T_1)+\pi(T_2)}$	Rate the stories based on how likely you and your allies are to win out over the sorcerer. If you expect your team to win, rate the story high. If you expect your team to lose, rate it low. Do not consider whether or not you *actually* win. Only rate the stories based on what you expected to happen *before someone gets defeated*.
directness	$\frac{\sum_{i=1}^{n} \text{closeness}_i(a_1,a_2)}{n}$	Rate the stories based on how close you are to the sorcerer. There are many kinds of closeness: physical closeness, emotional closeness, familial closeness, etc. Only consider the distance between *you* and the sorcerer.
intensity	$\max(\text{utility}(a_1,\emptyset),\text{utility}(a_1,T_1)) - \min(\text{utility}(a_1,\emptyset),\text{utility}(a_1,T_2))$	Rate the stories based on how much is at stake for you. Imagine how bad it will be if the sorcerer wins and how good it will be if you and your allies win. Stories which could end very badly or very well for you should be ranked high. Stories where your happiness is not likely to change very much should be ranked low. Do not consider the *actual* outcome of the story. Only rate the stories based on how much you think is at stake *before someone gets defeated*.
resolution	$\text{utility}(a_1,E) - \text{utility}(a_1,\emptyset)$	Rate the stories based on how much better off you are at the end. How much happier are you at the end of the story than at the beginning? Only consider how *you* have been affected. Do not consider how things *might have been*, only how they actually happened.

The study was conducted via a web interface in which participants could drag and drop stories from an initial random order into a sorted order of their choosing. Each participant ranked the same four stories for all four dimensions. Dimensions were presented to each participant in a random order.

All four stories had the same beginning, but different middles and ends. All stories were written in the second person such that the reader was the protagonist. All stories centered around a conflict between the reader and an evil sorcerer. This conflict with the sorcerer was the basis on which each story was ranked. The text of the stories was composed of simple actions which can be formally expressed as STRIPS-style planning operators (Fikes and Nilsson 1971). In other words, the stories were such that they could be produced by an automated planning system.

The content of the stories was structured so that, given our orderings for each dimension, no two stories would appear at the same index for the same dimension. That is, the second most intense story was never ranked second for any other dimension. Participants were not told of this constraint—it was imposed in an effort to minimize the chances that a participant would give the "correct" ranking based on a misunderstanding of the dimension's definition.

Each dimension is intended to be distinct from the others. A high value for one dimension should not imply anything about the values of other dimensions.

In order to avoid confusion from vocabulary, the dimensions were not given names in the study. Participants were simply given a description of the concept and asked to sort the stories. The descriptions of each dimension that were presented to the participants can be seen in Table 1.

We chose to create stories (rather than use excerpts from existing media) because it provided the opportunity to control for content, word choice, and length. As a result, these narratives are not "natural narratives," but ones contrived for this experiment in order to demonstrate specific qualities. One important direction for future work will be to test if the results of this experiment hold for stories that were not artificially designed.

Hypotheses

In this paper, we explore two main hypotheses:

1. For each dimension, participants will rank the stories in the order predicted by our formulas.
2. For each dimension, all participants will rank the stories in the same order.

The predicted orderings are given below for each dimension.

Note that this experiment does not require a commitment to specific formulas for $\pi(T)$ and $\text{utility}(a,T)$ as long as those formulas produce the predicted orderings.

For example, we assume that the knight is more likely to succeed when he has a sword and armor than when he has just a sword and no armor. It is not necessary to measure the exact difference in π between the two stories.

For each dimension, we provide a description of why the stories were ranked in their given order.

Balance Balance measures the relative likelihood of the reader and his allies to succeed. If the reader is likely to prevail, then balance is high. If the sorcerer is is more likely to prevail, then balance is low. We predicted this ordering for balance (from lowest to highest):

<table>
<tr><th colspan="2">Beginning</th></tr>
<tr><td colspan="2">This story takes place in a magical kingdom ruled by a wealthy king. The king has a young son, the prince. You are just a poor farmer, but you are friends with the prince. One day, an evil sorcerer kidnaps the prince! The king offers you a reward if you can get the prince home safely.</td></tr>
<tr><td><b>Story A</b>
You travel to the city. You ask a knight to kill the sorcerer. The knight buys a sharp sword at the market. The knight travels to the tower. The knight challenges the sorcerer to a fight to the death. The sorcerer reveals that he is your father. The knight defeats the sorcerer. The prince travels to the city. The king gives you a bag of gold. The king makes you a knight.</td><td><b>Story B</b>
You travel to the city. You ask a knight to kill the sorcerer. The knight buys a sharp sword at the market. The knight buys a suit of armor at the market. The knight travels to the tower. The knight challenges the sorcerer to a fight to the death. The sorcerer threatens to kill the prince. The knight defeats the sorcerer. The prince travels to the city. The king gives you a bag of gold.</td></tr>
<tr><td><b>Story C</b>
You travel to the tower. You challenge the sorcerer to a fight to the death. The sorcerer reveals that he is your father. The sorcerer threatens to kill the prince. You defeat the sorcerer. The prince travels to the city. You travel to the city.</td><td><b>Story D</b>
You travel to the city. You buy a sharp sword at the market. You travel to the tower. You challenge the sorcerer to a fight to the death. The sorcerer reveals that he is your father. The sorcerer and you become friends. The sorcerer defeats you.</td></tr>
</table>

Figure 1: The four stories used in the study. Each has the same beginning.

1. **C:** The protagonist (a poor farmer) fights the sorcerer with no equipment.
2. **D:** The protagonist fights the sorcerer after buying a sword.
3. **A:** The knight (acting on behalf of the protagonist) fights the sorcerer after buying a sword.
4. **B:** The knight fights the sorcerer after buying a sword and armor.

Directness Directness measures various kinds of closeness between the reader and the sorcerer. In stories A and B, the protagonist is interpersonally far from the sorcerer because a knight fights on his behalf. Our formulas predicted this ordering:

1. **B:** The protagonist and sorcerer are enemies, not related, and the knight fights for the protagonist.
2. **A:** The protagonist and sorcerer are enemies, family, and the knight fights for the protagonist.
3. **C:** The protagonist and sorcerer are enemies, family, and they fight face to face.
4. **D:** The protagonist and sorcerer are friends, family, and they fight face to face.

Intensity Intensity measures the stakes of the conflict with the sorcerer. Before starting the study, participants are asked to make two assumptions which are relevant to the utility function: it is better to be rich than poor, and the reader values his own life higher than the lives of other characters. Our formulas predict this ordering:

1. **A:** The life of the protagonist is not at stake because the knight fights the sorcerer. The life of the prince (a friend of the protagonist) is not at stake.
2. **B:** The prince's life is at stake.
3. **D:** The protagonist's life is at stake.
4. **C:** Both the protagonist's life and the prince's life are at stake.

Resolution Resolution measures the change in utility that the reader experiences relative to the beginning of the story. Our formulas predict this ordering:

1. **D:** The protagonist dies.
2. **C:** The protagonist succeeds but receives no reward.
3. **B:** The protagonist succeeds and is rewarded with money.
4. **A:** The protagonist succeeds and is rewarded with both money and knighthood.

Notes on Analysis

The data collected from each participant was an ordering of four stories for each dimension. The task of choosing an ordering is similar to classification, but it is important to note that two orderings can still be substantially similar even if they are not exactly identical. Capturing this degree of similarity is important, which precludes certain standard statistical tests.

For example, Cohen's κ coefficient is often used to measure inter-rater reliability, but κ assumes that the raters are choosing one of several discrete categories. The orderings (A B C D) and (A B D C) would be considered two different categories even though 5 of the 6 pairwise orderings are the same in both.

Another approach would be to consider the first, second, third, and fourth positions in the ordering to be categories. However, this enforces the constraint that, when comparing two orderings, if one element is in a different position then a second element must also be in a different position. According to this method, the orderings (A B C D) and (D A B C) are completely dissimilar despite the fact that 3 of the pairwise orderings are the same; in other words A comes before B in both, A comes before C in both, and B comes before C in both.

The edit distance metric, or Hamming distance (Hamming 1950), suffers a similar problem. The edit distance between two ordered sets of the same length is the number of substitutions that must be made in one set to transform it into the other. Using this metric, the distance between (A B C D) and (D A B C) is 4, the maximum possible.

Inversion Count

The study of sorting algorithms provides a useful metric for comparing two orderings: the number of inversions between them. An inversion is a pairwise difference between two

Table 2: This table shows the top 7 and the bottom 2 orderings for each dimension based the on average number of inversions from the orderings submitted by 30 participants. The orderings predicted by our formulas are highlighted in gray.

Balance		Directness		Intensity		Resolution	
Order	**Avg. Inv.**	**Order**	**Avg. Inv.**	**Order**	**Avg. Inv.**	**Order**	**Avg. Inv.**
C D A B	1.26667	B A C D	0.56667	B A C D	1.73333	D C B A	0.66667
C D B A	1.66667	B A D C	0.96667	B A D C	1.93333	D C A B	1.20000
D C A B	1.73333	A B C D	1.36667	A B C D	2.13333	C D B A	1.40000
C A D B	2.00000	B C A D	1.36667	B C A D	2.26667	D B C A	1.40000
D C B A	2.13333	A B D C	1.76667	A B D C	2.33333	C D A B	1.93333
C B D A	2.26667	B D A C	1.90000	B D A C	2.33333	D A C B	1.93333
D A C B	2.40000	B C D A	2.30000	B C D A	2.66667	D B A C	2.13333
(15 omitted)	(15 omitted)	(15 omitted)	(15 omitted)	(15 omitted)	(15 omitted)	(15 omitted)	(15 omitted)
A B D C	4.33333	C D A B	5.03333	C D A B	4.06667	B A C D	4.80000
B A D C	4.73333	D C A B	5.43333	D C A B	4.26667	A B C D	5.33333

ordered sets M and N. In other words, it is a pair of elements which appear in one order in M but in a different order in N.

Formally, let $\mathrm{index}(x, S) = 1$ just when x is the first element in ordered set S, $\mathrm{index}(x, S) = 2$ just when x is the second element in ordered set S, etc. Given two ordered sets M and N, an *inversion* is an ordered pair of elements (x, y) such that $\mathrm{index}(x, M) < \mathrm{index}(y, M)$ and $\mathrm{index}(x, N) > \mathrm{index}(y, N)$. This means that x is ordered before y in M, but x is ordered after y in N. The number of inversions between two orderings is a useful metric for analyzing this data because it captures the relative similarity of two orderings which may not be exactly the same.

When dealing with orderings of size 4, the minimum number of inversions is 0, meaning that both orderings are the same. There are 0 inversions between (A, B, C, D) and (A, B, C, D). The maximum number of inversions is 6, meaning that one is the reverse of the other. There are 6 inversions between (A, B, C, D) and (D, C, B, A), namely the pairs (A, B), (A, C), (A, D), (B, C), (B, D), and (C, D). All six of those pairs are ordered differently in the two sets. If we fix M and choose N at random, assuming that all 24 permutations of the 4 stories are equally likely, then there will be 3 inversions between the two orderings on average.

Results

30 people participated in the study—19 males and 11 females with a median age range of 26 to 35. No compensation was offered to the participants.

For a given dimension of conflict, let $\{P_1, P_2, ..., P_n\}$ be the orderings chosen by the n participants for that dimension (here, $n = 30$), and let $\mathrm{inversions}(M, P_i)$ be the number of inversions between two orderings M and P_i. For every possible ordering of the 4 stories M, we calculated its average number of inversions as:

$$\text{Avg. Inversions of } M = \frac{\sum_{i=1}^{n} \mathrm{inversions}(M, P_i)}{n}$$

To calculate the average inversions for the ordering $M =$(A B C D) for the dimension of *balance*, we calculate $\mathrm{inversions}((\text{A B C D}), P_i)$ for all 30 orderings P_i that were reported by the participants for *balance*; then we average those 30 values. An ordering's average inversions can be thought of as its average distance from each person's answer.

When some ordering's average inversions is low, that ordering is more popular—it agrees more with the orderings reported by participants. If all 30 participants had reported the exact same ordering, that ordering would be the most popular and it would have 0 average inversions. The reverse of that ordering would be the least popular ordering and would have 6 average inversions.

For each dimension of conflict, Table 2 presents the 7 orderings with the lowest average inversions (the top 7 best orderings for that dimension according to the participants). Table 2 also shows the two orderings with the highest average number of inversions (the 2 worst orderings according to participants) for each dimension.

Accuracy of Our Formulas

The orderings predicted by our formulas are highlighted in gray in Table 2. For the dimensions of *balance*, *directness*, and *resolution*, the ordering predicted by our formula has the lowest average inversions. For the dimension of *intensity*, the ordering predicated by our formula has the 5^{th} lowest average inversions.

This data supports our hypothesis that participants will rank stories in the same order as our formulas. Our formula for *intensity* may need to be improved based on these results to better agree with human perceptions.

However, demonstrating that our formulas can predict the best ordering is only helpful if there is a best ordering to be chosen. In other words, it is essential to demonstrate to what extent participants agree on a best ordering.

Participant Agreement

For this discussion, we will measure agreement with the "best" orderings as determined in Table 2 by the minimum average inversions. We are *not* measuring agreement with our formulas (though for 3 of the 4 dimensions, agreement

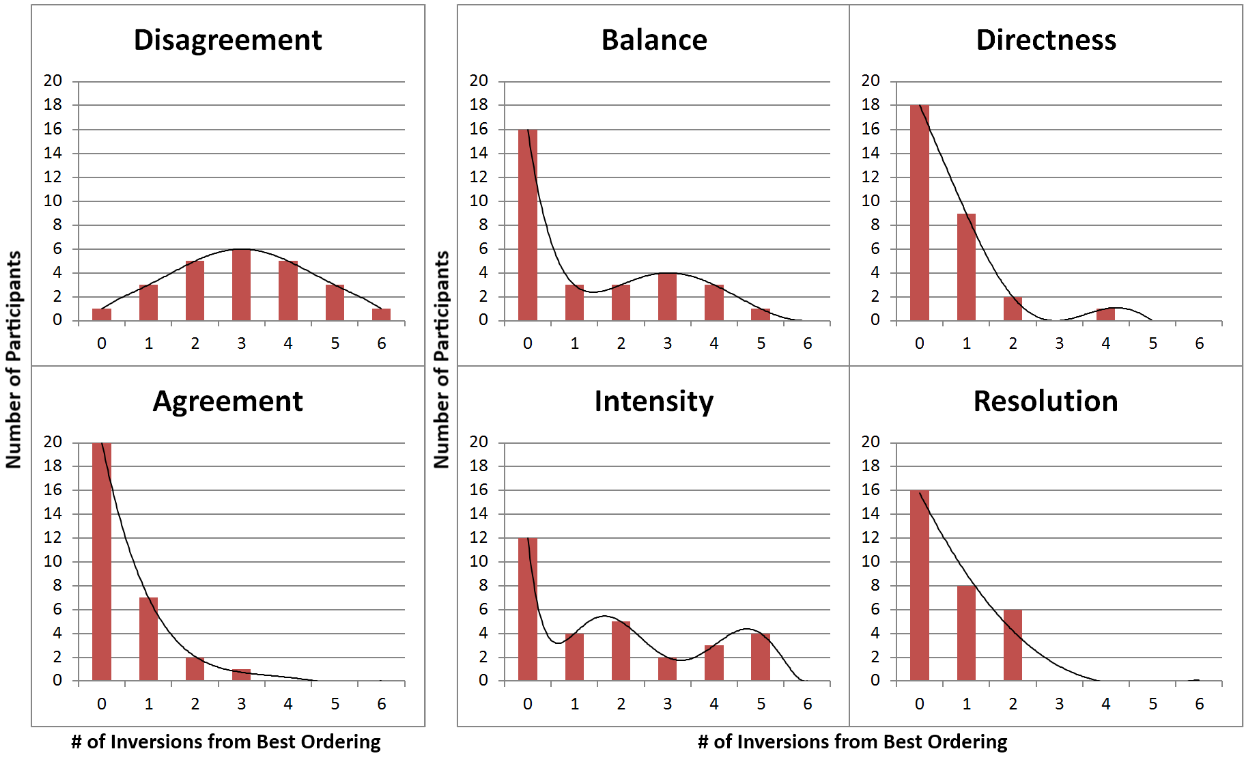

Figure 2: These histograms show how many participants (y axis) chose orderings that are some number of inversions (x axis) away from the best ordering for each dimension. Histograms representing disagreement and agreement are also provided for comparison.

with our formulas is equivalent to agreement with the best ordering). In other words, we wish to know how strongly the participants agree on the ordering they chose as best.

As discussed above, there is no clear way to calculate Cohen's κ coefficient to measure inter-rater agreement for this data. However, it is possible to express agreement by comparing our data to distributions representing agreement and disagreement using a goodness of fit test. Specifically, we use Fisher's exact test, rather than the traditional χ^2 test, because it is more accurate when comparing distributions which contain low expected values (Fleiss, Levin, and Paik 2003).

Measuring Disagreement Figure 2 plots the number of participants *y* who chose an ordering that is *x* number of inversions away from the best ordering for each dimension.

If there is complete disagreement on the best ordering, we would expect answers to appear as if they were given at random. This would result in a uniform distribution across the 24 possible permutations for the 4 stories. That uniform distribution, when plotted as the number of inversions from one specific ordering, is a roughly normal distribution (see *disagreement* in figure 2).

We predict, as a null hypothesis, that the observed distributions for each dimension will fit the distribution for disagreement. If that hypothesis can be rejected, we assume that some level of agreement was achieved.

Table 3 shows the p values returned by Fisher's exact test when comparing each dimension's distribution to the disagreement distribution. For all dimensions, $p < 0.05$, which is statistically significant. In other words, it is highly likely that the variance between our data and the disagreement distribution is due to something other than random chance. The null hypothesis is rejected.

Measuring Agreement Now we will explore the alternative hypothesis—that participants agree on a best ordering—in more depth.

If complete agreement had been achieved, all 30 participants would have chosen the same ordering, that ordering would be the best ordering, and it would be 0 inversions from itself, the best ordering. No dimension's data fits this "total agreement" distribution (see table 3), however Fisher's exact test and other similar tests give skewed results when applied to highly skewed distributions like this one for total agreement.

Given the subjective nature of how people perceive stories, it may be impossible to achieve total agreement. It is probably more helpful to compare against a distribution which indicates high (but not total) agreement.

One such distribution is given in figure 2. This distribution assumes that $\frac{2}{3}$ of the participants will choose the best ordering, and then the function will decay exponentially from there. When comparing our data to this agreement

Table 3: This table shows, for each dimension, the p values resulting from Fisher's exact test with the *disagreement, agreement*, and *total agreement* distributions shown in figure 2. We define our significance threshold to be $p < 0.05$. We observe that no dimension's data fits the disagreement distribution; only *intensity* does not fit the agreement distribution; and no dimension's data fits the total agreement distribution.

Dimension	Disagree	Agree	Total Agree
Balance	0.00290	0.14692	0.00002
Directness	0.00000	0.86897	0.00012
Intensity	0.02843	0.03312	0.00000
Resolution	0.00000	0.34509	0.00002

distribution, we hope to get *p* values which are *not significant*. In other words, it should be likely that the variance between our data and the agreement distribution is only due to random chance.

As indicated in table 3, the p values for the dimensions of *balance*, *directness*, and *resolution* were, as we hoped, not significant. This supports our hypothesis that participants agree on a best ordering for those dimensions.

The dimension of *intensity* fits neither the disagreement nor agreement distribution. We interpret this to mean that some level of agreement was achieved, but not to the same extent as for the other three dimensions.

Discussion

These initial results are promising, especially for *balance*, *directness*, and *resolution*. To the extent that we saw disagreement, we have identified several factors which may have contributed:

- Descriptions may not have been clear enough.
- Dimensions may have synergistic relationships.
- Participants were unable to ignore their knowledge of the story's ending.

Clarity of Descriptions Participants may have misunderstood the descriptions of one or more dimensions, which were intentionally brief and targeted at a high school reading level. We attempted to address this concern by running a small pilot study of the experiment before the version described in this paper. That pilot provided valuable feedback on how to clarify the definitions for each dimension. *Intensity* was the most widely misunderstood dimension in the pilot study, which may indicate why the data for intensity was the least supportive of our hypotheses.

It is also possible that participants misunderstood the events of the story. At least one participant indicated a misunderstanding of the outcome of story D. In an attempt to make the stories more "G rated," we used the text "X defeats Y" to describe the defeat action. This sentence does not make it explicit that Y is killed. Our predicted ordering for intensity is based on which characters' lives are at stake, so this may have caused confusion.

Dimension Synergy We assumed that each dimension could be measured independently of the others, but it is possible that participants perceived synergies between them. For example, if much was at stake (high intensity) but there was little chance that the sorcerer would prevail (low balance), participants might have given the story a low ranking for intensity. This might explain why story C is ordered before story D in the chosen best ordering for intensity. We hope to investigate how dimensions influence one another in future work.

Knowledge of the Ending The two dimensions with the least agreement—*balance* and *intensity*—require the reader to measure them independently of the actual outcome of the story. If the protagonist appears likely to prevail, balance should be high regardless of whether or not he or she actually does succeed. At least two participants had difficulty ignoring their knowledge of the actual outcome of the story. In future versions of this study, rather than ask participants to ignore the ending, we intend to leave the ending out.

Conclusions and Future Work

Previous research focused on developing a formal model of conflict that encompasses the entire phenomenon. This experiment was designed to validate four formulas for measuring specific dimensions of conflict which can be used to evaluate the content of individual stories. Based on our results, we draw three conclusions:

- The dimensions of *balance*, *directness*, *intensity*, and *resolution* are recognizable qualities of conflict.
- Human readers demonstrate considerable agreement on how to rank stories based on *balance*, *directness*, and *resolution*. Their rankings for *intensity* demonstrate less agreement, but are far from random and thus suggest that *intensity* is still a meaningful quality. We suspect that improvements to this experiment would yield higher agreement for this dimension.
- The orderings predicted by our formulas for *balance*, *directness*, *resolution*, and to a lesser extent *intensity*, corresponded with those chosen by human readers.

The higher goal of this research is to identify what measurable qualities of stories readers perceive and how they evaluate different stories based on those criteria. We believe that this research represents progress toward that goal because it identifies quantitative metrics for evaluating conflict.

In the future, we hope to improve our formulas based on this data, implement a system that produces stories based on our model of conflict, and guide the production of stories with constraints on the values of these dimensions. Constraints on each dimension will be based on observed patterns in various genres. For example, in most computer role playing games, the protagonist's conflicts with the antagonist become increasingly balanced and direct. Combined with the other three dimensions which are easier to measure—*participants*, *subject*, and *duration*—we hope to gain considerable control over the content and quality of the stories we produce.

Acknowledgments

We wish to thank the 30 anonymous people who participated in this experiment for their time and effort. We also thank the Liquid Narrative research group at North Carolina State University for their assistance in running a pilot study. Lastly, we wish acknowledge the National Science Foundation for its support of this research (award IIS-0915598).

References

Abbott, H. P. 2008. *The Cambridge introduction to narrative*. Cambridge University Press.

Barber, H., and Kudenko, D. 2007. Dynamic generation of dilemma-based interactive narratives. In *Proceedings of Artificial Intelligence and Interactive Digital Entertainment (AIIDE)*.

Egri, L. 1988. *The art of dramatic writing*. Wildside Press.

Fikes, R., 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.

Fleiss, J. L.; Levin, B.; and Paik, M. C. 2003. *Statistical Methods for Rates and Proportions*. John Wiley Sons, 3 edition.

Gerrig, R. J. 1993. *Experiencing narrative worlds: On the psychological activities of reading*. Yale University Press.

Hamming, R. W. 1950. Error detecting and error correcting codes. *Bell System Technical Journal* 29(2):147–160.

Herman, D.; Jahn, M.; and Ryan, M. L. 2005. *Routledge encyclopedia of narrative theory*. Routledge.

Peinado, F., and Gervás, P. 2006. Evaluation of automatic generation of basic stories. *New Generation Computing* 24(3):289–302.

Pérez y Pérez, R., and Sharples, M. 2001. MEXICA: A computer model of a cognitive account of creative writing. *Journal of Experimental & Theoretical Artificial Intelligence* 13(2):119–139.

Ryan, M. L. 1991. *Possible worlds, artificial intelligence, and narrative theory*. Indiana University Press.

Szilas, N. 2003. IDtension: a narrative engine for interactive drama. In *Proceedings of the International Conference on Technologies for Interactive Digital Storytelling and Entertainment*.

Vale, E. 1973. *The technique of screenplay writing*. Souvenir Press.

Ware, S. G., and Young, R. M. 2010. Modeling narrative conflict to generate interesting stories. In *Proceedings of Artificial Intelligence in Interactive Digital Entertainment (AIIDE)*.

Ware, S. G., and Young, R. M. 2011a. CPOCL: A Narrative Planner Supporting Conflict. In *Proceedings of Artificial Intelligence in Interactive Digital Entertainment (AIIDE)*.

Ware, S. G., and Young, R. M. 2011b. Toward a Computational Model of Narrative Conflict. Technical Report DGRC-2011-01, Digital Games Research Center, North Carolina State University, Raleigh, NC, USA. http://dgrc.ncsu.edu/pubs/dgrc-2011-01.pdf.