

NORTHWESTERN UNIVERSITY

Conversational Argument Strength and Burden of Proof

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL  
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Psychology

By

Jeremy N. Bailenson

EVANSTON, ILLINOIS

December 1999

© Copyright by Jeremy Bailenson 1999

All Rights Reserved.

## ABSTRACT

### Conversational Argument Strength and Burden of Proof

Jeremy N. Bailenson

This paper presents a model describing how people compute the strength of each speaker's position in a conversational argument. I utilize two measures of claim strength. Local strength measures the intrinsic contribution of a claim based on the explanation and evidence within the claim itself. Support strength measures the amount of support a claim provides as a whole (in relation to the other claims in the argument). The global strength model utilizes the strength of each claim and the relevance between claims to compute each speaker's contrast score. Judges determine contrasts for each individual claim by comparing it to the claim with which it clashes. Argument structure determines the clash relations between claims. Speakers gain the global strength advantage in disputes by accumulating more contrast than their opponents accumulate. The model fits available data and predicts burden of proof characteristics such as anti-primacy and effects of concessions and queries. Additional experiments test the model and compare it to alternatives. In some instances the alternative models that do not take into account argument structure fit more variance in the burden of proof data than does the contrast model. Implications and suggestions for new types of models are discussed.

### Author Notes

Thanks to my advisor Lance Rips and committee members Doug Medin and David Uttal. Also thanks to Michael Bailey, Sarah Brem, Norman Eliaser, Shira Gabriel, Pablo Gomez, Bradley Love, Elizabeth Lynch, the Medin Lab group, and William Revelle for helpful comments on earlier drafts of this paper, to Florence Sales for and unmeasurable gratitude to Mom and Dad.

Correspondence concerning this article should be addressed to Jeremy Bailenson, Department of Psychology, Northwestern University, 2029 Sheridan Rd., Evanston, Illinois 60208. Electronic mail may be sent via Internet to [jnb@nwu.edu](mailto:jnb@nwu.edu).

## Table of Contents

Introduction.....	1
The Uniqueness of Conversational Arguments.....	2
Burden of Proof.....	4
Factors that Affect Burden of Proof.....	10
Previous Models of Informal Argument.....	28
The Global Strength Model.....	45
Testing the Global Strength Model.....	55
Pilot Experiment 1.....	57
Pilot Experiment 2.....	64
Experiment 1.....	74
Experiment 2.....	87
Experiment 3.....	94
General Discussion.....	99
Tables.....	118
Figures.....	134
Notes.....	140
References.....	141
Appendices.....	150

## List of Tables

Table 1.....	118
Mean contextual strength ratings in Pilot Experiment 2 by position for speaker one and speaker two.	
Table 2.....	119
Standardized Means, correlations with burden of proof, standardized coefficients, and p-values for local strength ratings that are regressed on burden of proof for Pilot Experiment 2. The overall regression was not significant, $F(6,2)=3.13$ , $R^2=.90$ , $RMSD=1.55$ .	
Table 3.....	120
Standardized Means, correlations with burden of proof, standardized coefficients, and p-values for relevance ratings that are regressed on burden of proof for Pilot Experiment 2. The overall regression was not significant, $F(5,3)=.52$ , $R^2=.46$ , $RMSD=3.00$ .	
Table 4.....	121
Correlations with burden of proof, standardized coefficients, and p-values for contrasts that are regressed on burden of proof for Pilot Experiment 2. The overall regression was significant, $F(5,3)=8.69$ , $p<.05$ , $R^2=.94$ , $RMSD=1.04$ .	
Table 5.....	122
Standardized Means, correlations with burden of proof, standardized coefficients, and p-values for local strength ratings, prior opinion ratings, and outcome cost ratings that are regressed on burden of proof for Experiment 1. The overall regression was significant, $F(10, 21)=3.62$ , $p<.01$ , $R^2=.63$ , $RMSD=1.57$ .	
Table 6.....	123
Correlations with burden of proof, standardized coefficients, and p-values for local contrasts, prior opinion ratings, and outcome cost ratings that are regressed on burden of proof for Experiment 1. The overall regression was significant, $F(9,22)=4.99$ , $p<.001$ , $R^2=.67$ , $RMSD=1.45$ .	
Table 7.....	124
Standardized means, correlations with burden of proof, standardized coefficients, and p-values for support strength ratings, prior opinion, and outcome cost that are regressed on burden of proof for Experiment 1. The overall regression was marginally significant, $F(9,22)=6.87$ , $p<.001$ , $R^2=.74$ , $RMSD=1.29$ .	

Table 8.....	125
Correlations with burden of proof, standardized coefficients, and p-values for support contrasts, prior opinion ratings, and outcome cost ratings that are regressed on burden of proof for Experiment 1. The overall regression was significant, $F(9,22)=7.28$ , $p<.0001$ , $R^2=.75$ , $RMSD=1.26$ .	
Table 9.....	126
Table 9. Performance of the different models in fitting the burden of proof data by Experiment. The table includes $R^2$ and $RMSD$ values for regressions using the specified model to fit the burden of proof data. In each cell, $R^2$ values appear just above the corresponding $RMSD$ values. 1Br and 2Br signify one and two branch arguments, Add signifies Addendums, Conc signifies Concessions, All signifies a regression across experimental conditions, * indicates $p<.05$ , $^{\infty}$ indicates $p<.01$ , and $^s$ indicates $p<.001$ .	
Table	
10.....	127
Standardized Means, correlations with burden of proof, standardized coefficients, and p-values for local strength ratings, prior opinion ratings, and outcome cost ratings that are regressed on burden of proof for Experiment 2. The overall regression was significant, $F(10,22)=3.87$ , $p<.01$ , $R^2=.65$ , $RMSD=1.78$ .	
Table	
11.....	128
Correlations with burden of proof, standardized coefficients, and p-values for local contrasts, prior opinion ratings, and outcome cost ratings that are regressed on burden of proof for Experiment 2. The overall regression was significant, $F(9,22)=4.48$ , $p<.001$ , $R^2=.65$ , $RMSD=1.74$ .	
Table	
12.....	129
Standardized means, correlations with burden of proof, standardized coefficients, and p-values for support strength ratings, prior opinion, and outcome cost that are regressed on burden of proof for Experiment 2. The overall regression was marginally significant, $F(9,22)=4.62$ , $p<.001$ , $R^2=.65$ , $RMSD=1.73$ .	
Table	
13.....	130

Correlations with burden of proof, standardized coefficients, and p-values for support contrasts, prior opinion ratings, and outcome cost ratings that are regressed on burden of proof for Experiment 2. The overall regression was significant,  $F(9,22)=5.54$ ,  $p<.001$ ,  $R^2=.69$ ,  $\underline{RMSD}=1.63$ .

Table

14.....131

Percentage of burden of proof justifications by category for challenge structures and direct rebuttals in Experiment 3. Ap stands for anti-primacy, Str stands for claim strength, Sup stands for support, Agr stands for personal agreement, Evi stands for evidence, Rel stands for relevance, and Oth stands for other.

Table

15.....132

Normalized means, standardized coefficients, and p-values for one-sided support strength ratings, prior opinion, and outcome cost that are regressed on burden of proof for Experiment 1. The overall regression was marginally significant,  $F(8,23)=5.67$ ,  $p<.001$ ,  $R^2=.66$ ,  $RMSD=1.43$ .

Table

16.....133

Normalized coefficients and p-values for one-sided support contrasts, prior opinion ratings, and outcome cost ratings that are regressed on burden of proof for Experiment 1. The overall regression was significant,  $F(8,23)=3.80$ ,  $p<.01$ ,  $R^2=.57$ ,  $RMSD=1.62$ .

## List of Figures

Figure 1.....	134
A graphical representation of argument 1. Arrows indicate clash relations between claims. Note the beginning of the new branch in claim f which clashes not with the previous claim e but back with initial claim a in the dispute.	
Figure 2.....	135
A graphical representation of the effect challenges have on argument structure. Bold lines represent contrast relations. Notice how speaker B has an extra contrast in the Challenge Version, even though the two speakers offer the same number of claims in both versions.	
Figure 3.....	136
Results from Pilot Experiment 2. High values along the Y axis indicate high burden of proof for the first speaker while low values indicate high burden of proof for the second speaker.	
Figure 4.....	137
Average Local and Contextual Support Ratings from Experiment 1.	
Figure 5.....	138
Average Local and Contextual Support Ratings from Experiment 2.	
Figure 6.....	139
Average Local, Contextual Support, and One-Sided Support Ratings from Experiment 1.	

## INTRODUCTION

People have different opinions on a multitude of pressing issues that surround them; sometimes these opposing ideas clash. Disputes are a common occurrence in everyday discourse, be it among friends in informal peer settings (Resnick et al., 1993), family members (Dunn & Mann, 1987), children in an elementary school (Knudson, 1994), business associates during negotiations (O'Neill, 1991), students working together in order to solve problems (Trognon, 1993), or politicians discussing an issue (Lax & Sebenius, 1991). These arguments typically involve two or more individuals attempting to convince each other or a third party of a particular issue or claim, and can occur in a formal arena, such as in a planned debate, or in an informal manner, such as in discussions with relatives or colleagues. I do not intend the word "dispute" to include unproductive discourses laden with caustic remarks. Instead, my focus will be on more productive arguments in which the participants state their opinions and then appeal to reasons to support their positions. Research has shown that these interactions result in qualitative improvements in the participants' reasoning (Kuhn, 1997). The following paper will explore these conversational arguments--their nature, their structure, their participants, most importantly their assessment.

This paper is organized as follows. First, I discuss the ways that arguments developed within conversation are unique. Second, I turn to the notion of burden of proof, a concept arising predominantly from legal disputes. In the third section I describe

the factors judges rely on when assessing burden of proof, including the social aspects of the arguers, the structural configuration of the argument, the order of the specific claims in the structure, and the strength or convincingness of the individual claims. In the fourth section I review attempts by psychologists and other researchers to model argument processing and discuss ways those models could be augmented to account for burden of proof. In the fifth section I present a global strength model that computes an aggregate measure of claim strength for each speaker's position in order to predict burden of proof in conversational arguments. The model is based on the clash relations between the claims in the dialogue. Finally, I then describe a series of experiments that test the global strength model.

### The Uniqueness of Conversational Arguments

Often times persuasive messages take the form of an expository segment promoting a single side of an issue. These types of arguments are common in advertisements, political speeches, and in any instances where the arguer is trying to convince an audience of a position without the presence of opposition. Some studies have shown that when a speaker refers to an opposing position during these types of expositions, they often make their claims appear more even-handed (Voss & Means, 1991; Baron, 1991). Other research has demonstrated that acknowledging alternative positions can in fact have negative effects on one's argument by causing subjects to consider the opposing claims (Koehler, 1991, 1994; Kuhn et al., 1994). While this body of research sheds some light on the nature of expository arguments, referring to opposing

claims in these monologues does not change the fact that the argument is unilaterally presented--one speaker or organization presents their claims.

In fact, most previous experiments on arguments used unilateral descriptions, passages, or lists of traits as stimuli, and not an actual dialogue (see McGuire, 1985 for a review). There has been very little research dedicated to the process of reasoning during interactive arguments, as Trognon (1993) and Resnick et al. (1993) point out. Non-conversational passages tend to be quite different from everyday disputes that occur when two or more speakers interact and present their points of view at separate and alternating moments. During a dialogue, speakers have the option of making a number of acceptable moves (Orsolini, 1993). Each move can produce a unique result and change the direction of the conversation. According to Lewis (1979), participants keep track of the dialogue and employ differential strategies as it changes. New utterances affect the conversational record in a variety of ways, depending on the history of the conversation, the type of information assumed by the arguers, and the degree of cooperation between the speakers. In this sense, information in the form of a dialogue is subject to factors not present in passages by a single speaker.

Reichman-Adar (1984) also describes how dialogue-based arguments differ from other arguments. According to her theory, linguistic mechanisms and discourse rules govern the overall structure of the argument. During the dialogue, arguers use cue words and phrases (e.g., 'but first', 'by the way', 'anyway') in order to mold expectations concerning forthcoming claims and to interpret the inference relations between claims by

placing them in the context of an overall argument structure. In discourse based arguments, participants and judges extract information not only from the meanings of claims, but from the way those claims interact with opposing claims. Research that examines claims in isolation may be missing crucial parts of arguments--the ones inherent to conversational dynamics. One of these crucial factors is the notion of burden of proof, an intrinsic aspect of conversational arguments that will be discussed in the following section.

#### Burden of Proof

The participants of a dispute, as well as the observers, often make judgments as to which of the opposing sides is doing a better job arguing his or her position. Research on argumentation spans many disciplines, including psychology, philosophy, linguistics, and law. One notion embedded within this area of study is burden of proof, a concept developed by philosophers and lawyers.

When people in an argument make claims, they must defend those claims with accepted evidence. After making a series of statements, each speaker has some degree of burden of proof. In other words, the participants have to do a certain amount of persuading in order to prove they are correct. Assume two speakers, A and B, take opposing viewpoints in an argument. If speaker A presents claims that are far superior in strength than those of speaker B, then A would have little to do in order to prove his or her position. The burden would lie on B, who would then be forced to overcome the presumption created by the disparity between the quality of the two positions. In this

section, I first discuss the traditional notion of burden of proof used in formal judicial settings and then turn to examine burden of proof in informal arguments.

### Burden of Proof in Formal Arguments

In the United States judicial system, judges use rules to decide whether or not information is admissible as evidence, and jurors are asked to use them in order to assess the quality of the evidence (Wigmore, 1937). In formal legal arguments, rigid guidelines indicate presumption; one of the sides in the argument incurs some degree of burden of proof by default. In other words, even if the sides present arguments that are equally convincing and there is no way to resolve the dispute based on the evidence, a specified party still loses. The amount of default burden depends on the context of the argument.

For example, in a criminal trial the burden lies on the prosecution to prove beyond a reasonable doubt that the defendant is guilty. At the end of the trial, a judge is required to instruct the jury that the defendant is presumed innocent and that the burden of proof is on the prosecution, although there are certain cases where the burden actually falls on the defendant at the outset of a trial, such as in cases involving sexually abused children (Koszuth, 1991).

On the other hand, in a civil trial the burden of proof does not automatically fall as heavily on the prosecution. In the legal system, variation in burden depends on the ultimate cost (to the defendant) of the outcome of the argument (Farley & Freeman, 1995). The outcome cost of sending an individual to jail for life (the outcome of a criminal trial) is higher than the outcome cost of taking financial resources from that

individual (the outcome of a civil trial). Consequently, the burden for the prosecution is higher in the former instance.

Farley (1996) has created general proof levels that a claim must reach in order to overcome presumption and win an argument when that claim bears the burden. For a scintilla of evidence, there must exist at least a defensible argument supporting the claim, regardless of the possibility of defensible rebuttals to that claim. For a preponderance of evidence, there must exist more defensible arguments in support of the claim than in support of its negation. Finally, for dialectical validity, all arguments conflicting with at least one supporting argument for the claim must be defeated. In other words, this level of validity is the most stringent--even if the majority of arguments opposing a claim are defeated, so long as a single one of them remains the burden still lies against the claim. According to Farley, this differentiation parallels the use of burden of proof in the United States legal system, in that as the cost of being wrong in resolving the dispute gets higher, the burden of proof standards become stricter.

People do base their decisions on default burden of proof ; in a study by Kassin & Wrightsman, (1979), mock jurors were less likely to convict a defendant when they received the burden of proof instructions at the trial's outset than when they received the instructions at the trial's close. According to those authors, early presentation of the instructions affected how jurors interpreted the evidence more than the post-hoc presentation.

However, the precepts dictating proof are not as well defined in disputes that occur outside of a trial setting; participants of informal arguments in social situations are not held accountable to consistent instructions and rigid legal guidelines. Nonetheless, judges of informal disputes clearly assign burden of proof to one of the participants. The research described in the following section describes the general rules that people employ when assessing burden in these informal disputes.

### Burden of Proof in Informal Arguments

Burden of proof in informal arguments is unique. Unlike a formal argument, where the context rigidly dictates which side is assigned a higher burden from the outset, both speakers in an informal argument have a certain level of burden. According to Walton (1998), “any assertion by an arguer brings with it a burden of proof,” (p. 251). Specifically, Walton argues that public opinion determines the exact levels of burden for the assertion at the beginning of the dispute: “Generally, in persuasion dialogue, if a proposition is widely or generally accepted at a particular time, then there would be a presumption in favor of it so that anyone who challenges or critically questions it will have to meet this presumption with appropriately strong arguments that will be acceptable to doubters,” (p. 39). In general, when a speaker makes a claim, then it is his burden to defend that claim with evidence. Speakers that offer radical or unpopular claims have higher burden than speakers who present commonly accepted ideas.

Empirical research on informal disputes has shown that the first speaker to initiate a claim in the conversation has the greater burden (Rips; 1998, Bailenson & Rips, 1996).

Furthermore, according to Rips (1998), burden of proof in informal disputes increases for a speaker with a high number of claims that all the speakers in the argument reject (i.e., one speaker concedes that he was wrong about a claim that other parties reject) or that are still in contention. Along the same lines, Walton and Krabbe (1996) argue that burden of proof shifts in relation to the number of concessions made, such that a person increases his or her own burden of proof by conceding claims made by the opponent.

But arguers do not often explicitly concede points during the course of a debate, and it may be difficult in some instances to determine burden of proof levels by tallying concessions. Even when clearly losing the dispute, they provide justifications for their claims, which, according to Orsolini (1993), “make the speaker’s position less questionable by the recipient” (pg. 281). This underlying structure of argument in conversation (claims followed by justifications) is so fundamental that when absent (i.e., when a speaker provides a claim without a justification) other speakers in the dispute will point out the burden of proof by responding with a challenge to that initial claim (Garvey, 1987).

Work on informal arguments by Van Wallendael (1990) provides support for Walton’s notion that each speaker in an informal dispute has a specific burden of proof level. Van Wallendael has shown that subjects’ perception of burden of proof can be non-complementary, in the sense that evidence in favor of one position does not necessarily count against other viable alternative positions. According to her model, adjustments to competing hypotheses are more likely later in the reasoning process when

large amounts of information are available. Bailenson and Rips (1996) found similar results in a forced choice task. In that experiment, subjects saw three similar argument structures between the same two speakers. The content of the arguments was identical in all three structures, but each structure featured the claims in a different order. Subjects chose one of the three structures as the argument giving a specified speaker the most burden of proof. While subjects agreed that a certain argument structure gave one speaker (i.e., speaker 1) the highest burden, that same structure was not necessarily chosen least often for the other speaker (i.e., speaker 2).

In summary, even before an informal argument develops, the structure of the conversation and public acceptability of the claims often determine which speaker will have a higher burden of proof. Once the argument proceeds, speakers present evidence and use dialogue moves in attempting to reduce their burden. Even though there are no rigid guidelines for most types of disputes, audiences following informal conversations not only recognize the dialogue moves but come to expect them. A number of researchers have shown that when judges assess arguments, they rely on a number of strategies, including characteristics of the speakers of the argument, the overall structure of the argument, the order in which the claims are presented, and the strength or convincingness of the claims in the argument. In the following sections the manner in which each of these individual factors affect burden of proof will be described in detail.

## Factors that Affect Burden of Proof

### Characteristics of the Arguers

Work on argumentative discourse in social psychology has focused on the situational factors affecting persuasion and judgment (McGuire, 1985). Early research suggests that people are more likely to be swayed by those they perceive as experts than by those they perceive to be less qualified (Hovland & Weiss, 1951; Kelman & Hovland, 1953). In addition, the charisma of the participants often affects judgment. Subjects are more likely to agree with an attractive person's opinion than with an unattractive person's opinion (Chaiken, 1987). As a further example, Copper (1991) found that newscasters who smile are better able to convince the viewer to like an unpopular candidate than are newscasters who do not smile. In this sense, at the outset of the dispute, charismatic arguers would have a lower burden of proof level than would uncharismatic arguers who posit the exact same claim; consequently the charismatic arguer's claims would not need be as convincing as his opponent's. Furthermore, Walton (1998) indicates that burden levels vary with the public acceptability of the claim. Speakers who take advantage of these social factors may be able to reduce their burden by making unpopular claims appear acceptable.

In arguments occurring in discourse, the characteristics of the arguers themselves often play a large role in the structural development of the dispute. Resnick and colleagues (1993) monitored arguments among groups of college students debating an

issue. The researchers chose discussion participants such that each group featured members with prior opinions on both sides of the issues. During the conversations the researchers coded social dimensions such as non-verbal cues as well as the dialogue structure. According to these authors:

Several dimensions of the situation are likely to affect the course of reasoning. These include the relative social status of the participants, the goal adopted by the group for the conversation, the goals of each participant, the content of the conversation, the kind of physical displays (e.g., objects, pictures, graphs, and texts) available, who can see these displays, and who can manipulate them (pg. 348).

So, debates occurring in conversation often are subject to variable situational factors. The remainder of this paper will focus on the aspects of arguments specific to the dispute itself and will not explore these substantial social factors, even though judges certainly tend to be biased by them. In the experiments discussed in the current work, I attempted to hold these social factors constant across all conditions in order to ensure that they were not responsible for any burden of proof patterns.

### Argument Structure

In order to understand the nature of informal arguments, it is necessary to examine the content of the dispute, in addition to the traits of the speakers (Rips, 1998; Resnick et al., 1993). A typical dispute consists of a series of statements by the participants. For the purposes of this paper, I will call this collective group of statements

the argument and will refer to each of the speakers' explanatory or evidential statements as claims. Participants in an argument can contribute to the argument without making an evidential or explanatory claim. Queries (e.g., 'Why do you say that?' or 'What is your evidence for that claim?') are challenges to the opposing position, while concessions (e.g., 'Okay, you're right about that') acknowledge the truth of a claim or group of claims made by the opposition, propelling the conceded claim into a state of mutual acceptance, or common ground (Rips, 1998; Clark & Schaefer, 1989). All statements made by a given speaker constitute his or her position in the argument.

Sometimes a speaker summarizes his or her position in the first claim of the argument, setting the agenda for the dispute. According to Reichman-Adar (1984), the initial claims in the argument are the most influential, as they define the topic and the context for the dialogue. This notion is in line with earlier findings on text comprehension that demonstrate the importance of an initial topic sentence (e.g., Kieras, 1978). By introducing the argument with the first claim, the speaker takes a stand on a previously inactive issue, potentially beginning a dispute. Consider the following argument 1:

- (1) a. Elane: Capital punishment executions should be televised in order to deter future crimes.
- b. Billy: Executions should not be made public--that demolishes the privacy rights of the condemned.

c. Elane: When someone has committed a crime so heinous as to warrant an execution they lose all rights to privacy.

d. Billy: Just because someone is guilty of a certain crime does not mean they are not a human being.

e. Elane: Execution is reserved for situations where the most retributive punishment is in order--the criminal deserves no pity.

f. Billy: That may be true, but televised executions would not work as a deterrent anyway--capital offenders have no fear.

g. Elane: But when potential criminals actually see the punishment carried out they will be less likely to engage in criminal activity.

h. Billy: Criminals probably would be more likely to commit crimes if the executions were televised, just so their friends could see them on TV.

i. Elane: That's not true--most offenders rationally weigh costs and benefits before they commit crime.

Figure 1 shows the clash relations among the claims in the argument. A claim clashes with another claim if it specifically attacks that claim. In this example, Elane begins the dispute and sets the agenda by stating a concise summary of her position with claim a. The agenda being set, Billy clashes with her initial proposition by offering a rebutting statement (b) indicating his own position, namely that public executions are a bad idea. In this manner the argument progresses, each speaker presenting claims to counter preceding ones, an interlocking configuration subordinate to the context set by the initial

claim (a). However, notice Billy's maneuver with claim f. Here he offers a partial concession--not to Elane's position as a whole, but at least to claim e. After withdrawing his original line of reasoning from the context set with claims b-e (capital offenders have rights to privacy) he begins a new line of attack (publicized executions would not deter) which, instead of responding to the previous claim (e), retraces back to Elane's initial position and directly responds to her first claim (a). As a result, he begins a new line of attack and begins a separate branch in the argument, creating a two-pronged structure with separate hierarchies, both responsively subordinate to the context set by the first speaker's initial claim (Rips, 1998). In this manner, speakers use conversational procedures to assemble arguments to suit their needs.

Work done by Sillince (1995) provides a framework for these embedded, multi-branch structures. He calls this strategy of jumping from one local conclusion to another "changing focus". Many arguments are not linear, with each claim responding to the previous one; in fact many of them need to be followed with a "route map" where signs point out the relations between claims. For example, an analysis of the Watergate tapes (Linde & Goguen, 1978) shows many examples of these complex non-linear structures, composed by a number of foci within the argument scope. According to Sillince, the argument scope is the outer topical boundary of the debate set by the main conclusion. The argument focus is the particular stage reached at a particular juncture of a debate, wider than a single conclusion but subordinate to the general scope of the dispute. The focus can contain a number of propositions, both premises and conclusions.

Shift in focus occurs when one of the participants abandons the current conclusion and begins a new and relevant line of reasoning by offering another conclusion. Specific acts in conversations tend to produce this effect, for example, queries (i.e., 'What exactly are we arguing about?') or other utterances (i.e., 'But it still doesn't explain...', 'But it isn't THAT'). Focus shifts are often marked by specific frame words (Sinclair & Coulthard, 1975) or interruptions (Sillince, 1995), as the participants in the arguments change tactics and begin new lines of reasoning, and they are often used strategically by participants in a dispute.

The structure of the argument affects burden of proof in a number of ways. First, by examining structure it is possible to assess the impact of conversational moves such as queries, changing focus, and concessions. When a speaker begins a new branch by conceding a claim, they accrue more burden of proof (Walton & Krabbe, 1996; Rips, 1998). Furthermore, the structural impact of shifting focus and querying changes clash relations among claims, as Figure 1 illustrates. Later sections will examine the impact of these clash relations on burden of proof.

In summary, disputants can take specific actions in order to build the general architecture of the dispute. Within that architecture, speakers strategically arrange claims in a particular order. Previous research, most of which has regarded arguments outside of the conversational context, has examined some effects of this sequential ordering of claims in the argument. Examining the order of the claims without acknowledging their role in the argument's structure is not the optimal manner to analyze a conversational

dispute, but studies on sequential order of claims do provide some information as to how people process arguments.

#### Order of Propositions.

The manner in which arguers sequence their statements impacts their burden of proof. Presenting two groups of identical claims in different orders can produce two separate patterns of judgment as to how persuasive the claims are (McGuire, 1957) . For example, some studies find a recency effect, in that information at the end of an argument is retained better in memory and is more persuasive than information at the beginning of an argument. These effects have been found in situations where subjects read two contradictory messages (Crano, 1977) as well as in mock trials where subjects judged guilt or innocence (Furnham, 1986). Moreover, previous research on conversational disputes demonstrates that claims later in the dispute contributed more to burden of proof decisions than earlier claims (Bailenson, 1997).

Researchers also find ample instances of primacy effects in judgment. In these experiments, people retain information from the beginning of an argument more effectively than information from other serial positions, and this information then biases how later claims are processed. Pennington and Hastie (1986) showed this phenomenon in mock jurors: The first piece of information jurors received initiated a mental story representation of the supposed crime. In turn, that representation framed the interpretation of succeeding claims to make the information consistent with the overall story. Primacy effects also occur when judges over-estimate the frequency of instances

that occur early in a sequence and generalize the characteristics of those primary instances to the rest of the list. This type of primacy occurs in attributions of how other people perform a task (Benassi, 1982), in predictions of coin toss outcomes (Roth, 1975), and in judgments of contingencies (Yates & Curley, 1986). In addition, social psychology research on causality (Johnson et al., 1989; Vinokur & Ajzen, 1982) demonstrates causal primacy, the disproportionate weight given to events that occur at the beginning of a causal chain. Those researchers demonstrated that, regardless of the content of the proposition, initial events that occur before others are given more causal responsibility than later events in a causal chain.

In a review of the persuasion literature, Hogarth and Einhorn (1992) found systematic patterns of primacy and recency effects in various types of situations where subjects received information and formed judgments. They broke down tasks into categories based on information length, complexity, and time of judgment, i.e. whether subjects made the judgment after seeing all of the information (end-of-sequence), or whether they made it while processing each piece of information (step-by-step). For short, simple tasks they found recency effects for Step-by-Step response modes and primacy effects for End-of-Sequence. Time of judgment did not make a difference in studies using long and complex tasks. However, it is important to note that the studies Hogarth and Einhorn surveyed did not monitor when subjects actually encoded the information. So even in tasks where subjects made judgments at the end of the sequence, it is possible that they could have been updating their beliefs after seeing each new piece

of information. In other words, even though they were not told to respond until after seeing all of the information, they might have updated their impressions as they went along. Consequently the differences between the two presentation methods may be difficult to observe, since there is no way to monitor when the subjects actually formed their judgments. To summarize Hogarth and Einhorn's (1992) findings, the patterns governing primacy and recency, although they do exist, are quite complicated.

Previous studies of order effects typically interchanged the positions of claims or evidence with an unorganized serial list. This type of presentation (discrete pieces with little overall binding structure) finds its home predominantly in the laboratory. When people evaluate arguments in discourse, as discussed above, they cannot entirely dissociate the claims from the configuration of the dialogue as a whole. As a result, order effects found in a list of unbound claims may not hold when those claims are embedded in a conversational structure. Evidence for this hypothesis comes from a study by Bailenson and Rips (1996) in which subjects read naturalistic dialogues between two speakers having a dispute. The results demonstrated an anti-primacy effect, in that the first speaker who initiates the debate with a context-setting claim incurs more burden of proof than the second speaker, even though the individual claims of the two speakers are rated as equally strong when evaluated in isolation. This bias against the first speaker seems to be an effect of position that only surfaces in structured dialogue, as most previous studies have not shown a disadvantage in persuasion for items at the top of a

serial list. Consequently, it may be the case that the traditional notions of primacy and recency should be reevaluated in the context of natural interactive dialogues.

One plausible explanation for anti-primacy has to do with the roles of the two speakers. In work designed to study the manner in which people plan conversations, Scholtens (1991) describes the difference between the initiator of the conversation and the recipient<sup>1</sup>. The initiator “stages”<sup>2</sup> the argument, in that he or she makes the first claim that does not refute any existing claim. By setting the context, this first claim provides the underlying structure or script to be more or less followed for the remainder of the argument. According to Scholtens, the recipient does not have much control in molding the hierarchy of the conversation, as his job is primarily geared towards following the text and understanding: collecting information, signaling irrelevant ideas, pointing out falsities, and giving general reactions. On one hand, the analogy between the recipient and the second speaker of the dispute is incomplete, since both speakers can employ strategies throughout the argument to change the structure (i.e., Rips, 1998; Walton & Krabbe, 1996; Resnick et al., 1993). However, the division of roles is appealing since the first speaker occupies a structurally unique position; he or she does not respond to an existing claim. The model proposed in a later section posits that the different levels of burden of proof for the two speakers arise in part because of the difference between the types of conversational moves the two speakers offer.

Another shortfall of non-dialectic studies is that they don't address the full variety of conversational moves. For example, at any point in an argument speakers can

challenge each other with a query, such as “Why do you say that?” (Rips, 1998; Bailenson & Rips, 1996; Resnick et al., 1993). It is possible to study order effects concerning these types of actions, as speakers usually can invoke them at any point during the dispute. In a forced choice task, Bailenson and Rips (1996) showed that queries had more of an effect on burden of proof when they occurred late in an argument when the challenged speaker does not have an opportunity to respond. This finding exemplifies how the outcome of an argument can shift without adding or changing the order of any actual propositions.

In this section, I discussed studies examining the importance of order in assessing lists of information. While these studies provide insight as to how people process evidence, their relevance to burden of proof in conversational arguments is limited. In the following section, I present studies that describe how people determine the strength or convincingness of claims in an argument. Unfortunately, most of the previous research on claim strength is similar to the research on order effects, in that the studies examined the strength of isolated propositions, not the strength of statements in a structured dispute. Nonetheless, the work provides an excellent starting point for the goal of the current work: calculating conversational argument strength.

### Claim Strength

In order to reduce their burden of proof, speakers need to back up their claims with other claims that offer explanation and evidence. Certain types of claims in arguments are more persuasive than others, and consequently reduce burden of proof

more effectively than others. People pay attention to the strength of individual claims in an argument when they are actively involved in the outcome of the dispute, for example, when they are held accountable for their judgments (Tetlock, 1983) or when they have a vested interest or strong prior opinion in the issue being debated (Petty, Cacioppo, & Goldman, 1981). Strength in argument can be approached in a number of ways, but this paper draws the distinction between local and support strength.

By definition, local strength concerns the inherent characteristics of individual claims that make them convincing. Conceivably, the local strength of a claim is the same regardless of its context--whether isolated from the surrounding claims or placed in the argument context, the local strength should not change. Support strength is a measure of how well an individual claim reinforces the other claims made by a given speaker. In other words, a proposition has a high degree of support strength if it makes the argument in which it is housed more likely to be true (Rips, 1994; Osherson, et al., 1990; Collins & Michalski, 1989). Clearly the support strength of a speaker's argument is based somewhat on the local strength of the component claims—in order for a claim to reinforce the argument as a whole it has to present convincing evidence and explanations. However, the support strength also is based on how a claim relates to the other claims made by the speaker (as well as how it relates to opposing claims). The following section will further describe the difference between these two measures.

Local Strength. In the reasoning literature, the standards for assessing the strength of an individual claim vary. Early studies assumed that longer explanations

were stronger than shorter ones, and their strength manipulations were often based solely on length (e.g. Gulley & Berlo, 1956). However, it is easy to imagine ineffective long claims as well as persuasive concise claims. Consequently, most recent argument research has focused on other properties of claims.

One area of research examines the difference between offering actual evidence for a claim and offering an explanation. For example, work by Cathcart (1955) confirmed the intuition that judges prefer claims citing evidence over completely unsubstantiated claims. On the other hand, additional studies show that people are more persuaded by explanatory mechanistic information conveying a causal story than by statistical evidence merely indicating a correlation (Ahn & Bailenson, 1996; Ahn, Kalish, Medin, & Gelman, 1995; Kuhn, 1991). According to those researchers, strong claims are those presenting explanations in order to support a causal story to further the goals of the argument. These results are related to Kuhn's (1994) work that highlights subjects' inability to distinguish between explanation (a potential causal hypothesis) and actual evidence (the data to support the hypothesis). In her experiments, judges often gave more weight to mere explanations than such unsubstantiated hypotheses should warrant. However, work by Brem and Rips (in press) shows that subjects can provide and correctly evaluate actual evidence when they are not constrained by their own lack of data or when they are aware that evidence supporting a particular hypothesis exists. Their research describes protocol experiments in which, given proper definitions, subjects can in fact distinguish between the two types of statements.

Baron (1991) had subjects rate the relative strength of a number of different types of claims. He found that subjects perceived definitive statements showing conviction to be convincing; statements beginning with strong terms, such as "I am sure" or "without a doubt" showed conviction and consequently persuaded subjects. In addition, his subjects preferred claims that were "correct", in other words whose content was consistent with their own beliefs. So researchers have identified a number of characteristics that are responsible for determining local strength, including the size of the claim, the amount of evidence and explanation within the claim, and whether the content of the claim is consistent with the beliefs of the judge.

Support strength. In a structured dispute, claims contribute to support strength as a function of how well they fit into the overall structure of the argument and on how they relate to the other claims around them. In a later section, I describe models that use support strength as a way to model argumentation. In this section I focus on research that differentiates between propositions of high and low support strength.

Psychology research has dedicated a large deal of attention to support strength in inductive arguments that generalize a category (Osherson, et al., 1990; Sloman, 1993; Rips, 1989; Medin et al. 1997). For example, consider argument 2:

(2) A robin has sesamoid bones.

A turkey has sesamoid bones.

A parrot has sesamoid bones.

In this argument, the two premises lend support to the conclusion. Moreover, the amount of support each premise contributes depends on how it relates to the other premise (as well as on how it relates to the conclusion). For example, research by Osherson et al. (1990) demonstrates that the strength of this argument is proportional to the degree of coverage of the two premises. In other words, if the premises are diverse (i.e., they maximize the similarity to the lowest category that includes both the premise and conclusion categories) then the argument will be stronger. So the support strength of the second premise, "A turkey has sesamoid bones," depends on its similarity to the first premise, "A robin has sesamoid bones." In this sense, the support strength of a claim cannot be determined independently.<sup>3</sup>

Support strength also depends on the manner in which claims are arranged. Early research exploring the relation between claims of differential strength focused on the differences between the "climax" and "anti-climax" order (Phillips, 1908). The climax order features the strongest claim at the end of an argument, while the anti-climax order presents the strongest claim at the beginning. Sponberg (1946) found better retention for claims presented in the anti-climax order, in addition to a slight advantage in attitude change. However, other researchers have found the opposite effect, that the climax structure was more persuasive and maximized support (Cromwell, 1950; Gulley & Berlo, 1956).

Hogarth and Einhorn (1992) found claims presented in the strong-weak (anti-climax) order to be less persuasive than arguments presented weak-strong (climax), but

only in situations where subjects updated their beliefs as they encoded each new piece of information. According to them, a weak claim has less of an effect when presented initially because the judge has not yet anchored a strong position on the issue and he or she does not have an expected level of strength with which to compare the initial weak claim.

The studies discussed in the previous paragraphs in this section changed the order of isolated persuasive statements--not statements in a structured dialogue. Theories of conversation suggest that dialogues function correctly only when the speakers follow a mutually shared set of conventions (Van Eemeren et al., 1996; Grice, 1989; Jackson and Jacobs, 1980). As a judge processes an argument he or she expects a certain amount of relevance (Sperber & Wilson, 1986) between claims, in that a proposition needs to be responsive to the preceding propositions in order to provide support. In conversational debates, one way a speaker reduces burden of proof is by clashing with his opponent. In observing conversational argument strategy employed by debaters, Wood (1968) and Freeley (1966) indicate that refuting the opponent's claims with evidence and explanation is the basic way in which arguers support their position. This makes intuitive sense; in disputes where the arguers get to refute each other's claims in front of a judge, an individual statement only adds support to a speaker's position if it is more convincing than the opposing speaker's refutation of that claim.

Previous research shows that in arguments between multiple speakers, rebuttals serve an arguer predominantly when they respond to an existing claim (Koller, 1993).

Koller's studies show that rebuttals initiated alone have an extremely adverse effect on the speaker. In contrast, when the same rebuttals are executed in response to an existing claim, they provide more support. For example, in his third experiment, subjects read the accusation that "a visiting professor from India was allegedly working on experiments in order to manipulate gender in human embryos" (p. 383). Subjects were convinced more by the rebuttal, "he is not performing any gene research in order to manipulate gender" (p. 383) when it was preceded by the accusation than when it appeared alone.

The importance of clash in conversational arguments can also be seen when arguers present a type of claim called an undercutting defeater (Pollock, 1989). When a speaker offers an undercutting defeater, he does not challenge the truth of a proposition. Instead, he calls into question the degree of clash between two propositions (i.e., 'Your last claim has nothing to do with my argument'). Although previous research has shown that undercutting defeaters are not as common as rebutting defeaters (Shaw & Johnson-Laird, 1993), in a recent study Rips (1998) demonstrates that subjects who rated speakers' acceptance levels of individual claims are sensitive to differences between the two strategies. Furthermore, the importance of clash in the outcome of disputes has also been demonstrated in presidential debates (Morello, 1988).

To summarize the research on support strength, many factors contribute to the manner in which a claim provides reinforcement to others. In this section I discussed the amount of coverage among the claims by a speaker, the order of the claims, and the relevance of the claim to other claims in the argument. However, support strength is also

somewhat influenced by local strength, since a claim needs to be convincing in and of itself in order to provide reinforcement to other claims. Ideally, support strength would measure only the reinforcement a claim provides to the other claims made by that speaker, not local strength.

Previous studies on strength of argument have focused on local strength of individual claims and on how those claims interact with each other. However, few studies study local strength or support strength in two-sided, conversational arguments. Furthermore, there is a lack of research relating claim strength to the notion of burden of proof in argument. In the current work, I attempt to do just that.

### Conclusions.

There have been few studies that systematically examine conversations within argument. If a model is to accurately account for assessment burden of proof, then it must take into consideration the phenomena surrounding burden in informal conversational arguments discussed in this section: 1) independent burden of proof levels for each speaker, 2) anti-primacy (the high burden assigned to the first speaker), 3) order effects (specifically, claims late in the argument contributing most to burden of proof), 4) the impact of the variety of conversational moves (including queries, shifting focus, claim strength, and retraction), and 5) the importance of clash.

Previous models of argument processing have proceeded along two major lines. One class of representations emphasizes the degree of support that propositions within the dispute provide one another. The second class focuses on conversational aspects of

arguments. In the following section, I will discuss in detail specific models from these two perspectives and the ways that the models either account for or could be modified in order to account for the above-described burden of proof phenomena.

### Previous Models of Informal Argument

There has been a multitude of approaches to understanding how people construct, comprehend, and judge arguments. This section is divided into two major parts. The first discusses two models based on support relations between propositions in arguments. Earlier I discussed research that differentiated between high and low support strength. Now I turn to models that use the notion of support to represent argument, focusing on two seminal representations: Toulmin's (1958) theory, a classic model that represents informal one-sided arguments by demonstrating the specific types of support relationships, and Thagard's (1992) ECHO, a model that expands the notion of support to include arguments featuring both sides of an issue. The second part turns to models which, instead of focusing on support relations, center on the structure of the dialogue in which disputes appear. Specifically I discuss four different classes of models: Bayesian inference, Social Reasoning, Pragma-Dialectics, and Commitment. While the examples discussed here, as well as in previous sections (i.e., Reichman-Adar, 1984; Hogarth & Einhorn, 1992), by no means exhaust all the relevant argument models, they do highlight the major concepts to be used in the construction of the burden of proof model described later on in this paper.

### Models based on Support Relations

Models based on support relations emphasize the manner in which propositions provide reasons for each other. A proposition can provide a strong entailment relation, deductively ensuring that when the claim that provides support is true then the supported claim is also true. Rips (1994) and Johnson-Laird (1991) describe models that account for these types of deductive relations. However, other models focus upon support relations between claims that are inductively strong but not deductively valid (Collins & Michalski, 1989; Toulmin, Rieke & Janik, 1979; Toulmin, 1958). In the current section I will discuss two of these models, beginning with Toulmin's.

Toulmin's model. Toulmin (1958) describes a structural basis for informal arguments. According to Toulmin, when assessing an informal argument, a judge should rely on two kinds of considerations. The first kind depends on knowledge that is domain dependent; it is only relevant in limited areas of discourse (e.g., law or medicine). When evaluating an argument in a specific field, the judge can partly base his or her decision on the criteria applying to that field. However, there are structural aspects of arguments that are domain general. Examples of domain general components are "claims" (the conclusions speakers want the judge to accept), "grounds" (the data constituting the factual evidence in the argument), and "warrants" (the rules linking the grounds to the claims). Warrants can be either explicitly stated in the argument or implicitly based on prior knowledge. These parts of the argument provide a framework above and beyond

their evidential or explanatory content; they indicate the structural arrangement of the argument's statements.

Toulmin recommends that people who analyze disputes should attend to these field invariant factors, because an argument lacking any of the three is most likely weak. Even in situations where the judge lacks knowledge concerning the specific topic of discourse, he or she can consider these factors in order to make their evaluations. A strength of Toulmin's work is that it provides a tool to analyze arguments from an empirical standpoint. Psychologists (Voss, Blais, Means, Greene & Ahwesh, 1986) and legal analysts (Wangerin, 1993) have incorporated Toulmin's ideas into their models of reasoning. In addition, Toulmin's model provides a workable avenue for representing the way arguers handle burden of proof. According to Walton (1998), when speakers makes claims, each speaker automatically accrues burden of proof. The way they reduce that burden is by citing data, warrants and justifications--the concepts developed by Toulmin--to support that claim. In this sense, speakers who use these tools effectively will have less burden of proof than speakers who make claims without using them.

It is difficult, however, to apply the model to the other burden of proof phenomena. Toulmin's work is limited in that it does not address the dynamics of multi-sided arguments--the structural properties are, for the most part, designed to represent arguments in which the support and inferences all center upon a single position.

ECHO. Thagard (1992) offers an approach to understanding arguments similar to Toulmin's in that both models specify support relations between claims in an argument.

Thagard's model, however, takes this concept a step further and allows claims not only to lend support but to contradict each other, hence allowing for more elaborate arguments with multiple positions. The system decides between competing hypotheses by determining their degree of explanatory coherence based on the available evidence; people accept arguments that best explain the data. In ECHO, Thagard's connectionist implementation, claims are represented in a network of relations featuring excitatory links between propositions (and data) that cohere and also inhibitory links between propositions that rebut one another. The model provides a number of principles formalizing the notion of explanation and coherence, such as priority for data over explanations, preference for simple explanations over complex ones, and avoidance of contradictions.

Thagard (1989) applies his model to real world arguments, such as the debate between Creationism and Darwinism, as well as certain criminal proceedings. After receiving specific inhibitory and excitatory links, the model cycles and eventually predicts a winning position based on explanatory coherence. ECHO successfully evaluates a number of historical arguments and makes an important contribution to the study of argumentation by modeling processes that occur during the evaluation of multi-sided disputes. In addition, it is also consistent with other theories which posit that people assess the strength of a claim by evaluating explanations (Pennington & Hastie, 1993; Sloman, 1994).

One problem with using ECHO to assess burden of proof lies in determining the relations of coherence and incoherence. For example, each claim providing evidence in favor of Darwinism may support a number of similar claims and also may refute claims in favor of Creationism. Consider the evidential claim, “Animals have instincts,” (Thagard, 1989, p. 448). This claim could conceivably support both Darwinian (i.e., instincts are inherited traits) and Creationist (i.e., God gave the animals instincts) interpretations. In this sense, relations may be subtle, and judges may interpret the nature of support and contradiction differently. If the model were adapted to take into account structural cue words (Reichman-Adar, 1984) or the general architecture of the dialogue (Rips, 1998) then these relations might emerge more clearly.

Furthermore, while the network is sensitive to support between claims, it currently has no mechanism to represent the order effects and linguistic structural properties discussed in previous sections. The network could most likely be modified to weigh the disproportionate impact of recent claims on burden of proof. Moreover, if one hypothesis is determined to have a higher burden at the outset (i.e., the first speaker), then the network could be given a higher tolerance level for that hypothesis. The tolerance level is a mechanism within ECHO that specifies the amount of activation a particular hypothesis needs in order to rule out competing hypotheses. It might prove more difficult to represent the effect of conversational moves such as queries that do not provide explanation within the network. CONVINCER, the model to be discussed next is similar to ECHO in that it is capable of representing probabilistic support relations along

two sides of an issue. However, CONVINCENCE also takes into account the conversational aspects of the dialogue structure itself.

### Models Based on Dialogue Structure

This section discusses models that focus on dialogue-based disputes. Arguments in conversation are governed not only by support relations, but also by the types of dialogue moves that the speakers in the argument employ. Here, I describe four models that outline the types of moves that arguers make in a conversation and how those moves affect judges' perception of the dispute and of burden of proof. While there is other research that focuses on the structure of dialogue in argument (e.g. Muntigl & Turnbull, 1998; Jackson & Jacobs, 1980), I discuss the four specific representations that are most relevant to the new model I describe later in the paper. The section begins by outlining a Bayesian inference network called CONVINCENCE.

CONVINCENCE. Kim and Pearl (1987) describe CONVINCENCE (A Conversational Inference Consolidation Engine), a system that formally models reasoning as a Bayesian network of causal schemas. The system features nodes (variables or groups of propositions) linked together in causal connections. The representation is composed of target nodes (variables that directly affect the outcome of decisions), data nodes (observable variables), and intervening nodes that link data nodes to target nodes<sup>4</sup>. Each node can take on multiple roles in that a single node can be the cause of one set of variables while simultaneously being the manifestation of another set. The purpose of CONVINCENCE is to help the user of the system organize information about certain beliefs

and to condense an argument into a form that will improve the user's chance of attaining a quick, optimal decision about a probability. One of the most striking characteristics of the model is the procedure it uses in order to construct the representation--it relies on dialogue. By orchestrating a discussion between itself and the user, Kim and Pearl's model extracts the information most relevant to the argument and formulates a structural network.

Once the network is set up, the system employs a Bayesian inference procedure to derive the probabilities of the relevant variables. So, for example, consider a sample argument concerning the verdict of the O.J. Simpson civil suit. The system would ask the user to rate a series of probabilities concerning both competing hypotheses, such as "What is the probability that O.J. left his glove at the crime scene?" and "What is the probability that O.J. is guilty given that he left his glove at the crime scene?" Then, after assimilating all the probabilities, the system reports two (complementary) likelihoods, one for guilty and one for innocent. The system can also operate on non-binary decisions with more than two competing hypotheses.

CONVINCE aids the user in objectifying the structure and parameters of an argument, and works with him or her to attain a rational solution by having a conversation. However, the nature of the interaction is not one of clashing propositions and opposing viewpoints; on the contrary, the discussion is usually limited to the system asking the user to assign probabilities. In this sense, the model would need to be fundamentally modified in order to account for a number of the burden of proof

phenomena, such as clash and anti-primacy. In other words, the two speakers in the conversation are not trying to reduce their own separate level of burden of proof. The interview instead is geared towards determining probabilities of events. Moreover, the system reports complementary likelihoods for those events instead of computing a separate level of burden of proof for each position. The studies discussed previously indicate that burden of proof assignment is not necessarily complementary.

In addition, CONVINCENCE can only represent causal links. A more exact representation would include the ability to symbolize class membership and object properties, allowing the system to process a broader range of arguments, perhaps including disputes other than those which gauge the probability of an event's occurrence. Nonetheless, even though Kim and Pearl's system does not feature a dialogue based on clash, it strives to model reasoning within the medium where it usually occurs-- conversations. The following section discusses the Social Reasoning model, which focuses less on probabilities and more on the types of conversational relations among propositions in a conversation.

Social Reasoning. Resnick, Salmon, Zeitz, Haley Wathen, Holowchak (1993) describe a model that attempts to understand reasoning as a social practice. They represent interactive arguments in discourse as a series of inter-connected idea units, with links delineating different thematic connections. Arguers support their conclusions with either theoretical premises, general beliefs or values supposedly shared by most members of the participants' culture (e.g. 'being treated like a human being is generally a good

thing') or factual premises, causal links that relate a specific claim or position to a theoretical premise. An example of a factual premise is 'We should not execute criminals because it robs them of their humanity'. In other words, executing criminals violates the notion that, in general, people should be treated like human beings.

Theoretical premises may not appear in the actual dialogues, as by definition they are accepted by the arguers. As a result, they are not vulnerable to many conversational moves that factual ones are, such as challenges, answers to challenges (justifications), and concessions.

Resnick and colleagues provide a number of tools to analyze conversational arguments. Their representation features idea units linked thematically such that connections reflect both semantic support and refutation. Furthermore, idea units provide information regarding the structure of the dialogue, since the units convey the type of conversational move that occurred. For example, different units are used to represent elaboration of another's idea, an elaboration of one's own idea, a repetition of a previous statement, a challenge to another statement, or an answer to a challenge. In addition, the model incorporates social factors, such as turn changes and group cooperation.

Experimental support for the model comes from analysis of arguments made by triads of university students debating nuclear power policy. The researchers found that in their natural dialogues the students relied on the various conversational moves described in the model:

Participants listen carefully to each other and construct their arguments in relation to what others say. More specifically, they appear to build complex argument and attack structures. People appear to be capable of recognizing these structures and of effectively attacking their individual components as well as the argument as a whole. The discussants disarm anticipated (or previous) arguments first by making concessionary statements and then by weakening or attacking those statements (pg. 362).

One strength of their representation stems from being grounded in actual conversations between disputants. The focus of the system is representing different methods of clash, which is one of the fundamental contributors to burden of proof. Furthermore, the model is capable of depicting conversational moves such as queries and retractions. However, it lacks mechanisms to demonstrate how those moves affect burden of proof. In other words, Resnick et al. provide an account of how the argument proceeds, but not of how a judge perceives the outcome of the dispute. Without such a mechanism, it is difficult for the model to account for most of the phenomena that surround burden of proof, such as anti-primacy and order effects. However, the model does provide an excellent foundation that may somehow be expanded in order to account for burden of proof.

I now turn to another model capable of richly representing the structural aspects of a conversational dispute, the pragma-dialectic model. Unlike social reasoning, this

model is more concerned about how a productive dispute should proceed, as opposed to representing the exact manner in which people do tend to argue.

Pragma-Dialectics. Van Eemeren and Grootendorst (1984;1992) describe the pragma-dialectical theory of argumentation, one that integrates the normative rational aspects of argumentation with the descriptive discourse-based aspects of the conversation. According to these authors, a conversational argument should proceed along specific ideological starting points. First, speakers should publicly identify their commitments by stating them within the conversation. Second, the argument should be part of an interactive discourse procedure where one of the arguers takes on the role of the protagonist, the individual who defends a certain position, and the other acts as the antagonist, the person who attempts to refute that position. Finally, the participants should cooperate in resolving their difference of opinion and further the benefits of critical discussions by maintaining a relevant and rational exchange of ideas.

In the pragma-dialectical theory of arguments, a conversational dispute proceeds in four stages. In the confrontation stage, participants in the argument display a difference of opinion. In the opening stage, the protagonist and the antagonist identify themselves and their initial commitments. In the argumentation stage the antagonist methodically attacks the standpoint under discussion while the protagonist defends the standpoint. In the concluding stage, the disputants decide whether the protagonist's position has been successfully defended against the points brought up by the antagonist. If the parties do not both agree on the outcome, then the argument has not been resolved.

The model also includes a series of pragma-dialectical rules that should bind speakers in the conversational dispute, including rules governing when parties can offer claims, what kind of claims they can offer, and what other conversational moves (such as requests for justifications and retractions). Furthermore, Van Eemeren, Grootendorst, and Snoeck Henkemans (1996) define burden of proof as follows: “A party that advances a standpoint is obliged to defend it if asked by the other party to do so” (pg. 283). However, they also indicate situations where (without violating the rules of the dispute) the protagonist evades this assignment of burden of proof, such as when that speaker presents his or her initial claim as self evident, or by giving a personal guarantee of the correctness of the standpoint. Note that this condition is similar to Walton’s (1998) claim that burden of proof varies as a function of public approval of the claim.

The strengths of the model lie in the principles it provides for individuals who need to resolve a difference of opinion through dialogue. Furthermore, the representation provides mechanisms to account for the burden of proof phenomena that occur in conversational arguments. Given the differentiation between protagonist and antagonist, the model can accommodate anti-primacy based on those disparate roles. In addition, the order effects could be represented as a function of the particular stage in which a proposition occurs. For example, recent claims (especially concessions and summaries at the end of the argument) may impact burden of proof most because the claims made in the concluding stage are more central to the argument’s resolution than claims made in other stages.

However it is important to note that this model is only effective as a procedural aid to arguers who seek the truth; it provides a single mechanism to judge the outcome of a dispute, since both participants have to agree on whether or not the original standpoint offered by the protagonist is valid. While this mutual agreement may occur in some types of disputes, in others such as presidential debates and trial situations, both sides of an argument tend to stick firmly to their position. These latter types of disputes are the ones which the pragma-dialectical model has trouble assessing.

Commitment. Walton and Krabbe (1995) present a rule-based model that centers on the notion of commitment to explain two-person, argumentative dialogues. Their model is an extension of Hamblin's (1970) work and focuses on arguments that occur in everyday conversations. According to Walton and Krabbe, many types of arguments exist in dialogue, including formal dialogues, persuasive dialogues, and negotiations. For the most part, the authors focus on persuasive dialogues, in which participants who initially possess conflicting points of view try to resolve their conflict by persuading each other of their position.

Walton and Krabbe claim that the main determinants of how the argument proceeds in this type of dialogue are the commitments made by each party, both the explicit commitments that surface in the form of assertions (light side), and the implicit ones contingent upon internal beliefs (dark side). Parties are bound by their prior commitments, especially the internal dark side ones. The extent of these commitments determine a distinct burden of proof level for each participant in the dispute. The authors

notably go into some depth when formalizing the rules of what they call the "Permissive Persuasion Dialogue.

Walton and Krabbe are somewhat vague as to why certain rules are intrinsic for reducing burden of proof in an informal argumentative dialogue. According to those authors, in certain types of informal dialogues, victory by one party only occurs by an opponent's retraction. However, pilot data from an experiment where subjects constructed their own arguments shows that retraction almost never occurs even though subjects clearly could point to a winner in the dispute. In a more general setting, Walton (1988) notes that burden of proof in persuasion dialogues can depend on one party proving that a position "is at least plausible on a balance of considerations" (p. 38). Nonetheless, the model could more accurately represent burden of proof by taking into account different degrees to which claims are convincing, instead of determining the contribution of each claim solely by its structural function; previous research has shown that subjects rely heavily on claim strength when assessing burden of proof (Bailenson & Rips, 1996). Finally, the model would benefit from empirical support that tested whether its mechanisms would demonstrate other burden of proof phenomena, such as anti-primacy and order effects.

Rips (1998) provides empirical support for a theory in which commitments in arguments are structured in terms of the conversational moves in the dispute. Statements relate to one another not only through inductive and deductive support, but also in ways that depend on the structure of the conversation. Based on different types of

conversational moves such as making claims, challenging with a query, or providing an answer to a challenge, the model monitors the level of commitment of each speaker at each point in the dialogue. For example, when arguers assert claims, they clearly commit to those claims. Likewise, when an arguer explicitly concedes a claim (as in claim f in argument 1), then he or she commits to the opponent's last claim. However, the arguer does not need to concede a claim to agree with an opposing claim; instead he may indirectly accept the opposing claim by agreeing to one of its justifications.

Rips's structural model predicts the way subjects assign commitment levels at various points in the argument. Speakers' commitment levels will vary according to the type of moves they employ in the structure. For example, subjects who judge arguments are sensitive to different types of claims, such as the difference between rebutting defeaters that directly refute a previous claim and undercutting defeaters that attack the relation between a claim and its support. Furthermore, participants did not interpret failing to respond to a claim as a concession. Consequently, unless a speaker explicitly concedes a claim they remain committed to that claim.

One of the advantages of framing an argument in terms of the commitments of each speaker is that burden of proof (and its surrounding phenomena) depends on commitment. Rips presents data that show how speakers committed to mutually accepted claims, ones believed to be true by both speakers, have lower burden of proof:

If a claim becomes mutually accepted, then the participant who offered it need no longer defend it. Since there are fewer claims that the participant

must now establish, his or her burden will be lighter. By contrast, a claim that is mutually rejected or open (i.e., not mutually determined), can increase a participant's burden. The participant must find new ways to support claims that are currently under contention and new ways of shoring up previous points when currently supporting claims become mutually rejected (p. 432).

Consequently, each speaker has a separate level of burden depending upon the number of his claims that have become mutually accepted. In addition, data from Rips' studies demonstrate anti-primacy, "presumably because the first claim takes a stand on an issue that sets the agenda for the rest of the discussion," (p. 433).

While clearly the structural relations among claims contribute to how people assess burden of proof in conversational arguments, as discussed above, a model that represents burden of proof must account for the strength of the claims on each side of the dispute. Rips's model provides the framework for a systematic representation of argument processing. However, the framework would benefit from an additional module that compiled the strength of all the claims offered by each speaker. One of the goals of this current work is to provide that module. In other words, by using the tools developed by Rips for organizing the structure of a conversational argument, the current model assesses the perceived overall strength of each arguer's position.

Summary. This section described a number of approaches to argument processing and how those approaches account for the phenomena that occur when judges

compute burden of proof. Some of the models are concerned with support relations between claims (e.g., Toulmin's model, Echo) while others are concerned with the structure of the dialogue (e.g. Convince, Social Reasoning, Pragma-Dialectics). Commitment models, which are concerned with both dialogue structure and (to some degree) support relations between claims, account for more burden of proof phenomena than do the other types of models. However, commitment models could be improved by taking into account the strength of the claims in the argument as well as the structural role of the claims.

In the following section I propose a new model that accounts for how people integrate the strength of all the individual claims made by a speaker into a global measure. First I describe the model, and then report experiments that test various aspects of the representation.

## THE GLOBAL STRENGTH MODEL

The current work specifies a model that demonstrates how judges of conversational arguments determine the overall strength of each speaker's position. Earlier in this paper, I described two ways to measure the strength of an individual claim. The first was local strength, the independent contribution of the claim. Local strength depends on the evidence and on the explanations offered within the claim itself and is not influenced at all by surrounding arguments. The second measure is support strength. A claim provides support for an argument by providing evidence and explanations for the other claims in the dispute and for the speaker's overall position. Note that these two measures are not independent. Having a strong local strength also helps a claim provide support for other propositions. In this sense, the support strength of a claim partially subsumes local strength.

In a given argument, the two speakers make a number of claims. Each claim has some degree of local strength and support strength. When assessing the dispute, a judge of the argument needs to assimilate the strengths of these claims into some type of aggregate position for each speaker. In this section, I describe a model that compiles claim strength. The model can be applied to either local strength ratings or to support strength ratings.

The model was designed to highlight clash relations. As discussed in the Introduction, arguments in conversation are different from traditionally studied, one-sided arguments in that they feature interaction between speakers. When a speaker in a conversational dispute offers a proposition, he is often responding to an existing proposition made by the other speaker. Consequently, judges of the dispute do not assess claims in a vacuum. Instead, they compare a given claim to the claim to which it responds.

In order to illuminate the interactive nature of conversational arguments, the model computes a contrast score, a measure of how strong the claims from one speaker are when compared to the surrounding claims of the other speaker. Contrast score combines a number of factors previously shown to impact burden of proof, such as the strength of individual claims, the relations between claims, the roles of each speaker in the dispute, and the commitment levels of each speaker. In general, however, a contrast score simply measures how strong a speaker's position is in reference to that speaker's opponent's position. In this section of the paper I focus on describing the manner in which judges compute contrast scores and also the manner in which contrast affects burden of proof.

#### Contrast in One-Branch Arguments

Initially, consider arguments where the speakers each are committed to the same number of claims in the dispute. For each claim, judges compute a contrast by comparing the claim's strength to the strength of the claim to which it responds (i.e., with

which it clashes). Furthermore, each claim is assigned a relevance score. The relevance score is similar to Walton's (1998) probative relevance, where "a proposition is probatively relevant to another proposition only if it can be used to prove or cast doubt on this other proposition according to the methods of proving or casting doubt appropriate for a type of dialogue," (p. 64). In this sense, the relevance score reflects the contextual and semantic relationship between claims. A score high in relevance is extremely responsive to the previous claim; the content of the claim in question directly addresses the specific proposition expressed in the previous one. Another way to describe relevance is that (assuming a higher strength than the previous one) a relevant claim would effectively refute the previous claim. Drake (1993) utilizes a similar distinction between what he describes as relevance and truth (convincingness). In a series of studies he shows differences in hemispheric activation between the two measures.

In one-branch arguments (i.e., arguments that do not feature concessions or focus shifts), claims always respond to the immediately preceding claim, and individual contrasts are computed as follows:

$$D_n = R_n \frac{S_n}{S_n + S_{n-1}}$$

where  $D_n$  is the contrast of claim  $n$  ( $n > 1$ ),  $S_n$  is the strength (either the local or support) of the claim being assessed,  $S_{n-1}$  is the strength of claim  $n-1$ , and  $R_n$  is the relevance of claim  $n$  to claim  $n-1$ . The model forms a ratio in order to represent the strength of one claim in relation to another. A ratio is advantageous because it automatically normalizes

the difference between clashing claims. Other models utilize similar functions when representing discrimination from memory (White & Wixted, 1999), auditory differentiation (Allen & Neely, 1997), and distance comparisons (Birnbaum, Anderson, & Hynan, 1989).

The relevance score comes from subjects ratings of each pair of claims and is normalized to reflect each subjects' maximum use of the rating scale<sup>5</sup> (e.g., Schonemann & Lazarte, 1987):

$$R_n = \frac{O_n}{M}$$

$R_n$  is the relevance score,  $O_n$  is the observed rating by the subject along a scale, and  $M$  is the maximum rating that the individual subject utilized on the scale.

Note that it is possible for a claim to clash with more than a single claim in the argument. In these instances, support strength may be a better measure than local strength, since the former takes into account all the previous claims in the argument. Nonetheless, the purpose of designing the model such that a claim clashes with a single opponent claim is to attain a quantifiable representation of how judges compare opposing propositions by two arguers.

The first claim in an argument necessarily has no contrast since it does not clash with any existing claim. Punishing the first speaker for making a claim that doesn't respond to an existing claim is one of the ways in which the model expresses the anti-

primacy effect. Note that this does not mean the first claim has no importance in the dispute; the first claim still affects the value of the contrast for the second claim.

Assuming that speakers alternate turns, the contrast score is computed by comparing contrasts of the two speakers as follows:

$$G = \sum_{i=2}^n -1^{(i+1)} D_i$$

where  $G$  is the contrast score of the argument,  $D_i$  is the contrast of a specific claim  $I$ , and  $n$  is the total number of claims in the argument. According to the formula, the contrasts from one speaker are subtracted from the contrasts of the other. When a judge assesses burden of proof, she is guided by the sign of the contrast score. A positive score reflects an advantage for the first speaker, while a negative score reflects an advantage for the second speaker.

Anti-primacy (in arguments where the two speakers offer an equal number of claims) often arises because the second speaker gains the contrast score advantage by accruing one more contrast than the first speaker. When the first speaker makes the initial claim, that claim does not refute any explicit claim and as a result does not contribute a contrast. In arguments where the first speaker presents one more claim than the second speaker, the model would not necessarily predict anti-primacy since each speaker adds a comparable number of contrasts to the global score.

#### Contrast in Multi-branch arguments

As Figure 1 demonstrates, in some arguments a speaker can temporarily or permanently halt a line of reasoning and attack a prior claim or a position from a different

angle by starting a new branch. Consider the dialogue from argument 1. Define topic claim as the first claim of the new branch. In argument 1, this would be claim f. Define reference claim as the claim to which a topic claim clashes and refers to, often the first claim of the argument (see the link between claim a and claim f in Figure 1). To compute the contrast for a topic claim, the strength of that claim is compared to that of the reference claim, not to the preceding one (which is the last claim of the old branch).

$$D_t = R_t \frac{S_t}{S_t + S_r}$$

Here,  $D_t$  is the contrast of a topic claim,  $S_r$  is the strength of the reference claim,  $S_t$  is the strength of the topic claim, and  $R_t$  is the relevance of the topic claim to the reference claim. Then the score is computed down the branch as in any other string of refuting claims and continues until the argument ends or a speaker begins the next branch.

### Concessions.

The above section describes a process that applies when an arguer begins a branch while keeping the previous one open, that is, without conceding his or her claims from the previous branch. However, it is possible to drop the assumption that the speakers in the argument are committed only to their own claims. When an arguer concedes an opponent's claim or group of claims, those claims become mutually accepted. Correspondingly, claims the conceder made earlier that clashed with the now conceded claims become mutually rejected. As a result, concessions increase the amount

of the conceder's claims that get mutually rejected and increase the amount of claims by his or her opponent that become mutually accepted. Previous research has shown that if one speaker has a higher ratio of mutually accepted claims to mutually rejected claims than the other, then that speaker has less burden of proof (Rips, 1998). In this situation, a speaker does not add contrasts to his or her score from any claims that are mutually rejected.

The model determines mutual acceptance and rejection in accordance with the principles of commitment outlined by Rips (1998). Data from Rips's experiments show that when a speaker explicitly concedes a branch in a two-branch argument, most subjects perceive that the speaker still remains committed to most of their own previous claims in that branch. However, according to his rebutting principle (a), "A participant who accepts a rebutting defeater to a claim rejects that claim, unless the defeater is undercut for the participant" (p. 420). In the contrast model, a speaker does not gain a contrast from that mutually rejected claim. Consider argument 1. In this argument, when Billy offers claim f, he concedes that claim e is true. Consequently, claim e becomes mutually accepted. However, claim e is a rebutting defeater to claim d. According to rebutting principle a, by accepting that defeater, Billy rejects his own claim d. Since it is mutually rejected, Billy does not receive a contrast for that claim. However, In Rips's studies, most subjects did not implement this principle in its unrestricted form—they perceived that Billy was still committed to claim d. In the current studies I fit the model with and without rebutting principle (a) and compare the results.

### Challenges.

A challenge (e.g., ‘Why did you say that?’) is a form of clash; even though a speaker who queries her opponent does not offer any evidence or explanation to refute the previous claim she still alters the course of the conversation. One functional aspect of utilizing a challenge during a conversational turn is propelling the oncoming clash between claims to a later point in the argument sequence, as the scenario depicted in Figure 2 illustrates.

If speaker X has a responsive claim, B, in mind when his opponent, speaker Y, offers claim A, that speaker can present B or can save it, instead opting to challenge claim A by saying: “What evidence do you have for that?” At this point Y must further defend A by offering A<sub>1</sub>. A<sub>1</sub> will usually be a qualification, elucidation, or an expansion of A. Speaker X now can either accept the argument presented in A or alternatively can present claim B (or a version of B that is modified in response to A<sub>1</sub>). By making a challenge, X makes his contrast at slot 4 in the conversational sequence instead of at slot 2.

In this sense, a challenge serves to put the status of the challenged claim on hold. According to the model, in the challenge structure neither A<sub>1</sub> nor the challenge adds to the contrast score. Neither of them provides contrast since they are not substantive claims that clash with another. While A<sub>1</sub> might provide some support for the speaker’s overall position, it does not directly respond to an opponent’s claim. Therefore, the

model, which was designed to highlight clash relations, does not give Speaker 1 a contrast for claim A<sub>1</sub>.

If each speaker has the same number of claims, making a challenge in an argument robs your opponent of a contrast, as the only clash in the challenge condition occurs when claim B responds to A<sub>1</sub>. In this sense the challenge is a form of attack. In Figure 2, bold arrows indicate the points of contrast. Note that in this dialogue, speaker Y could offer an additional claim that refutes claim B. However this would require an additional speaking turn, and often times the context of the conversation may put constraints on the number of utterances each speaker can make.

#### Limitations of the Model.

The current model was designed to apply to specific types of arguments. Later in the General Discussion, I will discuss in detail the strengths and limitations of the model. However, in this section I briefly outline the types of disputes for which the model works best. Initially, the contrast score only takes into account a claim's relation to the immediately preceding claim by an opponent. In this sense, it applies to informal disputes that feature arguers alternating speaking turns. In other types of arguments, this type of a representation may have difficulties. Consider a formal presidential debate, where each speaker gets allotted a certain amount of time. Within that time period they make a series of claims. Their opponent then has a certain amount of time to respond by offering a group of clashing claims. In this sense, the conversational structure does not clearly dictate clash relations between claims, since there is no opportunity for opponents

to respond to claims as they are made. While the model could be amended to represent these types of disputes, in the current form it cannot.

The model makes an assumption that judges of arguments process and pay attention to every single claim. In other words, every claim receives an independent contrast. Consequently, the model applies best to situations where the judge is paying a high amount of attention to the dispute. In arenas where the judge does not have a reason to follow the dialogue diligently, other methods of aggregating strength may be more appropriate.

Finally, as previously discussed, the model does not apply to one-sided arguments. Moreover, the model does not apply to arguments that consist of a simple exchange of ideas. Consider a conversation that consists of two propositions. Speaker 1 offers claim A, and then speaker 2 offers claim B. If the exchange ends at this point, then the model does not effectively represent the situation, since speaker 1 will never get a contrast and cannot win the argument (regardless of the relative strengths of the two claims). The model could be amended to apply to these simple exchanges of viewpoints by giving the first speaker a modified contrast (perhaps by comparing the first claim to prior opinion for the topic). However, in its current form the global strength model was designed to represent discussions where the participants present their viewpoints as well as provide reasons and justifications in support of those viewpoints.

## TESTING THE GLOBAL STRENGTH MODEL

In this section I test the model. The purpose of the experiments is to demonstrate that 1) speakers reduce their burden of proof by gaining contrast in argument, and 2) aggregate claim strength cannot be determined independently of argument structure. Both of these results would provide direct support for the Global Strength Model.

I test the model in two ways. First, I report analyses that test the model's predictions on data from pilot studies. These studies focus on the contrast score parameter based on local strength ratings and they compare that score to burden of proof ratings. In the second part of this section, I design a series of new experiments in which it is possible to formally test both the local strength contrasts and the support strength contrasts. Moreover, in the new studies, subject groups were surveyed about their opinions concerning the topics of arguments. When testing the model I can take into account those prior beliefs. In both the pilot studies and the new studies, the model makes qualitative predictions about the data.

In all the studies reported, I used the same general methodology. Subjects read conversational arguments between two characters; they assessed the strength (or relevance) of the individual claims and they picked which speaker had more burden of proof. I then compared the burden of proof data to the patterns that the Global Strength Model predicts. The section begins by reporting the two pilot studies.

In Pilot Experiment 1, subjects examined one- and two-branch arguments. The arguments could either feature the same number of claims by each speaker or feature an extra claim by the initiator. The model predicts anti-primacy for arguments where the two speakers present an equal number of claims but not for arguments where the initiator has an extra claim, since in this latter situation the speakers each contribute the same number of contrasts. In the two-branch arguments used in the experiment, the second speaker always conceded the final claim of the first branch when he or she shifted to the new branch. As a result, it is possible to test the concession rules of the model outlined above. The speaker who concedes a branch loses at least one of his or her previous contrasts; as a result anti-primacy should be reduced when the second speaker concedes since he or she advances fewer contrasts.

In Pilot Experiment 2, I formally tested the contrast score parameter of the model by having separate groups of subjects rate the local convincingness of claims in isolation, the relevance of each claim, and the burden of proof score for each argument. Subjects read one-branch arguments composed of six claims. The model predicts anti-primacy, since speaker one has one fewer contrast contributing to the score than does speaker two. In addition, the model predicts that contrasts late in the model will contribute more to burden of proof than earlier contrasts, for reasons outlined above. It is possible to compare the model's results with baseline models consisting of local strength scores computed not by contrasts but by merely taking the difference or ratio between the sums

of the local strength of the claims from each position. The model should significantly outperform the baselines.

### Pilot Experiment 1

#### Method

Materials. I constructed booklets of 16 different argument topics consisting of either nine or ten statements apiece. Each argument appeared on a separate page. The content of the arguments varied, but typically involved entertainment, politics or college life, as argument 3 illustrates:

- (3) a. Calvin: I think that more than just Democrats and Republicans should be allowed to participate in presidential debates.
- b. Ronnie: There is no point in listening to people's views who have absolutely no chance of winning the election.
- c. Calvin: Third parties have a chance to win--look at how well Ross Perot did in 1992.
- d. Ronnie: In the past two decades, no third party candidate has captured more than a quarter of the electoral votes--it would be a waste of time.
- e. Calvin: Well maybe if they got opportunities to voice their opinions in public forums they would do better.
- f. Ronnie: That may be true, but third party candidates are all a bunch of crackpots anyway.

g. Calvin: Just because they are not in the mainstream doesn't mean they are not serious.

h. Ronnie: Third party candidates advocate all sorts of zany ideas, banning government, legalizing crime, and all sorts of things.

i. Calvin: Many of the third party ideas are valid solutions--ones that help the environment and alleviate social ills.

j. Ronnie: There is no way to know whether or not their solutions are valid because they have never been in power.

Claim j is underlined to indicate its absence in the condition where the initiator has more claims than the recipient. Note that the argument is in two branches; the recipient (Ronnie) begins a new line of reasoning with a concession when she presents claim f. The second branch (3rd party candidates are crackpots) is still relevant to the main agenda of the argument (pro/con of 3rd party candidates in debates) but is not entirely dependent on the claims presented in the first branch (they have no chance of winning). All of the 16 argument topics appeared in both one-branch and two-branch structures. In this experiment, the recipient always began the second branch with a concession. Here is an example of the same argument in a one-branch structure:

(4) a. Calvin: I think that more than just Democrats and Republicans should be allowed to participate in presidential debates.

b. Ronnie: There is no point in listening to people's views who have absolutely no chance of winning the election.

c. Calvin: Third parties have a chance to win--look at how well Ross Perot did in 1992.

d. Ronnie: In the past two decades, no third party candidate has captured more than a quarter of the electoral votes--it would be a waste of time.

e. Calvin: Well maybe if they got opportunities to voice their opinions in public forums they would do better.

f. Ronnie: Public forums wouldn't help--third party candidates are all a bunch of crackpots anyway.

g. Calvin: Just because they are not in the mainstream doesn't mean they are not serious.

h. Ronnie: Third party candidates advocate all sorts of zany ideas, banning government, legalizing crime, and all sorts of things.

i. Calvin: Many of the third party ideas are valid solutions--ones that help the environment and alleviate social ills.

j. Ronnie: There is no way to know whether or not their solutions are valid because they have never been in power.

The only difference between the one- and two-branch structure lies in claim f, in that speaker 2 either continues the line of reasoning begun in claim b (without using a conversational cue to signal a new branch) or offers a partial concession to claim e and begins a new line. The two branches were designed to be similar enough to be connected

into a single line of reasoning but sufficiently distinct to be seen as alternative points when separated.

Design. There were two groups of participants, one of which saw arguments with nine claims (i.e. the initiator had an extra claim) and one which saw arguments with ten claims (the two speakers had the same number of claims). For each group, participants received two booklets with every one of the 16 argument topics. In one of the booklets, they were asked to choose which speaker had more burden of proof, and in the other they were asked to rate the strength of each claim. The order of the booklets was counterbalanced such that half of the subjects in each condition rated the strength of the claims first and half of the subjects chose which speaker had more burden of proof first. Each of the two booklets contained exactly the same arguments (same versions of all of the 16 topics) in the same order.

Each subject received eight one-branch and eight two-branch arguments. For each topic a subject only saw one of the versions (i.e., subjects never saw the same argument topic twice with a different number of branches or a different number of claims). Every subject saw the topics in one of 12 different random orders. Across subjects, each version of each topic appeared an equal number of times.

Procedure. The first page of the burden of proof booklet contained instructions and a sample argument. Subjects read the instructions on their own. They were told they would see a series of arguments and were to decide “which of the two people has got more work to do in order to prove that they are correct.” This specific wording was

designed to prevent subjects from associating the term “burden of proof” with courtroom procedure, and has been used in previous studies on burden of proof (Rips, 1998; Bailenson & Rips, 1996). They were instructed to circle the name of the person they felt had to do more to prove that they are correct. Also, they were asked to rate how confident they were in their choice, by circling a number on a scale from one to seven, one being extremely low confidence and seven being extremely high confidence.

The first page of the strength booklet contained a sample argument and instructions on how to rate convincingness. Subjects rated each of the claims in all arguments on a seven-point scale, with one corresponding to an unconvincing claim and seven corresponding a very convincing claim.

Subjects all received the burden of proof booklet and the strength rating booklets (not necessarily in that order). No subject began the second booklet until everyone in the session was finished with the first one. Sessions typically took about an hour.

Participants. Fifty-two subjects participated in the experiment in order to fulfill a requirement in an introductory psychology course. Twenty-eight subjects assessed the nine-claim arguments and 24 subjects assessed the ten-claim arguments. Subjects were all native speakers of English, and were tested in small groups of up to six people.

### Results and Discussion

For each subject I computed a burden of proof score by multiplying his or her confidence ratings (from one to seven) by his or her speaker choice (1 for first speaker, -1 for second speaker). The grand burden of proof mean was  $-.35$ , indicating that there was

a slight bias overall to choose the second speaker more often than the first speaker.

There was a main effect for number of claims in that there was anti-primacy for ten-line arguments ( $\underline{M} = .21$ ,  $\underline{SD} = 4.51$ ) but not for nine-line arguments ( $\underline{M} = -.94$ ,  $\underline{SD} = 4.79$ ),  $\underline{F}(1,50)=11.62$ ,  $p<.01$ . There was also a main effect of branch, in that the first speaker had less burden of proof in two-branch arguments where the second speaker conceded ( $\underline{M} = -.69$ ,  $\underline{SD} = 4.70$ ) than in one-branch ( $\underline{M} = -.04$ ,  $\underline{SD} = 4.67$ ),  $\underline{F}(1,50)=4.43$ ,  $p<.05$ . The interaction between branch and number of claims was significant,  $\underline{F}(1, 50)=4.49$ ,  $p<.05$ .

Figure 3 shows the means of nine and ten claimed arguments by branch. The only condition in which anti-primacy occurs is one-branch, ten-claimed arguments. This pattern is generally consistent with the contrast model, since the only condition in which the contrast score favors the second speaker is when that speaker makes more contrasts than the first speaker. In the two-branch arguments, the second speaker concedes a contrast and loses his advantage, and in the nine-claimed arguments the speakers have an equal number of contrasts.

One problem with the data is the shallow slope of the line for two-branch arguments in Figure 3. In other words, the condition with the lowest mean burden of proof should be the two-branch, nine-claimed arguments. In this condition the initiator had the most contrasts (relative to the recipient), but the mean burden of proof is not extremely low. This may be due to a floor effect with the two-branch arguments. The mean of the nine-claimed arguments is slightly (but not significantly) lower than the ten-

claimed arguments, but since the two-branch arguments in general were low, the difference is negligible.

It is important to note that the lack of anti-primacy in the nine-line arguments is not merely due to having both the first and last claim in the argument, as previous work by Bailenson and Rips (1996) has shown that even when the initiator makes the final claim he still has the burden of proof (under conditions where the total number of contrasts were held constant).

It is not possible to fit the model formally to this data for the following reasons. First, crucial ratings are missing, such as the relevance scores of the individual claims. Furthermore, the local convincingness scores were not made in isolation; subjects rated each claim while it was in the argument structure. In order to avoid the influence of argument structure and relevance, there needs to be a separate group of subjects who rate local strength of the claims when separated from the argument structure. Pilot Experiment 2 and Experiments 1 and 2 remedy these shortcomings. However, while there is no mathematical fit to the data, Pilot Experiment 1 shows qualitatively that burden of proof does vary as a function of contrasts; in arguments that feature extra contrasts by the initiator, anti-primacy is reduced.

In Pilot Experiment 2, separate subjects rated local strength in isolation, the relevance score of each claim, and the burden of proof for one-branch arguments. I predicted anti-primacy in the six-line arguments, since the first speaker makes one fewer

contrast. Furthermore, it is possible to test the efficacy of the local contrast score in comparison to the baselines described above.

## Pilot Experiment 2

### Method

Materials. I constructed booklets of nine different arguments consisting of six statements apiece. Each argument appeared on a separate page. The arguments' content was similar to Pilot Experiment 1, as argument 5 illustrates.

- (5) a. Pat: Baseball has more breaks in the action than other sports.  
 b. Jim: Baseball has fewer breaks in the action than other sports.  
 c. Pat: You have to sit through all of the side changes.  
 d. Jim: The only long break comes after the seventh inning.  
 e. Pat: Every time they substitute pitchers there is another fifteen minute break.  
 f. Jim: At least there are not many substitutions in baseball compared to other sports.

Note that the argument is between two speakers who alternate when making claims. In order to ensure that the first and second speakers presented an equally convincing position, each argument appeared in two possible orders, such that each of the two characters in the argument presented the first claim (they were the initiators) in one of the two orders. I produced the alternative order of each argument by reversing the positions of claims a and b (initial claims), claims c and d (middle claims), and claims e and f (final

claims). So, in the above argument between Pat and Jim, in order to make Jim the first speaker, the orders of the claims appeared as follows:

- (6) a. Jim: Baseball has fewer breaks in the action than other sports.  
b. Pat: Baseball has more breaks in the action than other sports.  
c. Jim: The only long break comes after the seventh inning.  
d. Pat: You have to sit through all of the side changes.  
e. Jim: At least there are not many substitutions in baseball compared to other sports.  
f. Pat: Every time they substitute pitchers there is another fifteen minute break.

As a result, there were 18 different arguments, two different orders for each of the nine different topics.

In order to demonstrate the impact of conversational structure on burden of proof, it was important to ensure that the claims made by each speaker in the argument were roughly equivalent in convincingness. In other words, the total local strength of the three claims made by Jim in the above argument should be equal to the total strength of the three claims made by Pat. Without strong disparities in total claim strength in the arguments, subjects are forced to attend to argument structures. To this end, after creating the arguments, I ran a pretest in order to determine the strength of the claims made by each speaker, which will be described in a later section.

Design. Each participant received two booklets, each one with every one of the nine argument topics. In one of the booklets, they were asked to choose which speaker had more burden of proof, and in the other they were asked to rate either the local strength or the relevance of each claim. The order of the booklets was counterbalanced, so that half of the subjects rated the strength (or relevance) of the claims first and half chose which speaker had more burden of proof first. Each of the two booklets contained exactly the same arguments (same versions of all of the nine topics) in the same order. Consequently, each subject rated claim strength and burden of proof for the same nine arguments. For each topic they only saw one of the claim orders (i.e. subjects never saw the same argument topic twice with different speakers making the initial claim). Every subject saw the topics in a different random order. Across subjects, each of the two different claim orders for each topic appeared an equal number of times.

Procedure. The first page of the burden of proof booklet contained instructions exactly like those from Pilot Experiment 1. The first page of the local strength booklet instructed subjects to read each claim and decide “how convincing is each claim in and of itself?” Subjects rated each of the six claims in the argument on a seven-point scale, with one corresponding to an unconvincing claim, and a seven corresponding to a very convincing claim. The relevance packet asked subjects to decide “how well does each claim respond to the previous claim”, with a one corresponding to not responsive and a seven corresponding to very responsive.

Subjects all received the burden of proof booklet and either the strength or relevance booklet (not necessarily in that order). No subject began the second booklet until everyone in the session was finished with the first one. Sessions typically took about twenty minutes.

Participants. Forty-eight subjects participated in the experiment in order to fulfill a requirement in an introductory psychology course. All subjects were native speakers of English. Subjects were tested in small groups of up to ten people.

Pretest. In the pretest, subjects rated isolated claims in order to ensure that for all the arguments one speaker's claims were not more convincing than the other speaker's claims, since one of the goals of the study was to isolate effects due to the structure of the argument. Moreover, since the total strength for each position was the same, it should be possible to highlight the effect of local contrasts without them being masked by largely asymmetrical positions. To this end, after constructing the initial arguments I took the claims out of the argument structures and arranged them in a random order such that each one was isolated from its original conversation. In other words, there was a single randomized list with all the claims from the nine argument topics. In addition, I inserted "fillers", or slight variations of the original claims, in which the claims were more or less convincing. These fillers could be substituted for the original claims in order to adjust the strength of a given speaker's total position.

The instructions to the pretest were similar to the local strength condition in the main experiment. Thirty-two subjects were told to judge how convincing each claim was

on a seven point scale. The claims were presented in four different random orders. Participants were introductory psychology students, and sessions took about 20 minutes.

I averaged the strength of the claims offered by each speaker (three claims for two speakers in each argument). In six of the nine arguments it was necessary to use filler claims in order to equate the strength of the positions offered by the two speakers. After these substitutions were made, the average difference in the summed local strength of the two positions was .16 and the median difference was .15 on a scale from one to seven. The largest difference was .32,  $t(31)=1.45$ ,  $p>.05$ . By constructing the two positions in each argument to be equally convincing (in terms of the sum of the local strength of the component claims) the pretest data ensured that any differences found in burden of proof could be attributed to argument structure. It is possible to compare the pretest data to the contextual claim ratings from the main experiment. This data analysis appears in a previous work (Bailenson, 1997), and shows that the two measures are largely similar. The main difference between the two methods is that the initial claims (for both speakers) were rated weaker in the argument context than when they were rated in isolation.

### Results and Discussion

Across conditions, subjects chose the first speaker as having more burden of proof (anti-primacy) in 58% of the arguments. This difference was reliably different from chance,  $t(47)=3.28$ ,  $p<.05$ . Table 1 shows average local strength ratings of claims that subjects rated in the argument structure (i.e., not in isolation). There was no effect on claim strength for speaker, in that the convincingness ratings for any given claim was the

same regardless of whether made by speaker one or by speaker two. Local strength ratings varied according to the position of claims,  $F(2, 47)=88.77$ ,  $p<.01$ . Initial claims were rated as weakest ( $M = 2.67$ ), followed by middle claims ( $M = 4.47$ ). Subjects rated the final claims the highest ( $M = 4.76$ ). The same pattern occurred for normalized relevance scores, with initial claims were rated as weakest ( $M = .38$ ), followed by middle claims ( $M = .68$ ). Subjects rated the final claims the highest ( $M = .70$ ),  $F(2, 47)=54.09$ ,  $p<.01$

The contrast scores were calculated as follows. For each of the nine separate argument topics, I averaged across subjects and computed three measures: the isolated local convincingness ratings of each of the six claims (using data from the pretest), the relevance of the second through the sixth claims, and the burden of proof score for each of the nine topics. The contrast scores were computed according to the formulas for one-branch arguments depicted in the Model section. Then, burden of proof score (percentage of subjects who chose the first speaker) was regressed on contrast score. Contrast score was a significant predictor of burden of proof,  $F(1,7)=9.07$ ,  $p<.05$ ,  $R^2=.56$ ,  $RMSD = 1.79$ .

The contrast score was then compared to two baselines. The first baseline was a local strength ratio, computed by dividing the sum of the local strength ratings of speaker one's claims by the sum of the local strength ratings of the claims in the argument. The local strength ratio was not significant,  $F(1,7)=4.72$ ,  $p>.05$ ,  $R^2=.40$ ,  $RMSD = 1.98$ . The second baseline was a local-relevance ratio. This model was the same as the other

baseline, except instead of summing local strength ratings in the numerator and the denominator, it summed the product of each local strength rating and its respective relevance rating. In other words, it took into account both local strength and relevance, but didn't directly compare clash relations in a local ratio. This model actually outperformed both the contrast model and the baseline,  $F(1,7)=16.02$ ,  $p<.01$ ,  $R^2=.70$ ,  $\text{RMSD} = 1.48$ .

In sum, the baseline without relevance ratings did relatively poorly when fitting the data; the contrast model with the relevance ratings performed better. However, the baseline that took into account relevance ratings and local strength but did not compute actual local contrasts fit the data better than any other model. The models with the relevance ratings take into account the structure of the conversation--consequently, they should outperform the baseline without relevance. However, the more general model that matched relevance ratings with local strength ratings for each speaker and took a ratio performs the best. Apparently, enough clash information surfaces when a claim is matched with a relevance rating--computing local clash ratios may actually counterproductive. This hypothesis will be tested further in later studies.

In the next analysis, I regressed burden of proof scores on the individual local strength ratings of all six claims in the argument. Table 2 shows the standardized means, correlations with burden of proof, standardized coefficients, and p-values for each claim. Note that the coefficients for even claims should be positive while the coefficients for odd claims should be negative. I performed the same analysis using the relevance ratings

of claims two through six. Table 3 shows the results. Again, coefficients for even claims should be positive while the coefficients for odd claims should be negative. Finally, I regressed burden of proof on the average contrasts for each of the five serial positions. Table 4 shows the data. Coefficients for even claims should be positive while coefficients for odd claims should be negative. Even though the contrast regression had one less parameter than the analysis using the average ratings of the six claims, it had the best fit and was the only significant analysis,  $F(5,3) = 8.69$ ,  $p < .05$ ,  $R^2 = .94$ ,  $\text{RMSD} = 1.04$ . In addition, a recency effect was observed in that the only contrast that was a significant predictor was the last one between claim five and claim six (the coefficient for this claim was in the predicted direction, with a high contrast reducing the second speaker's burden of proof). Note that this analysis vastly outperforms the aggregate analyses in part because of the large number of parameters used to fit the data.

In summary, contrast score outperforms the baseline model without the relevance ratings. However, the aggregate model that fits the data best is the model with the summed products (of the local strength of claims and their respective relevance ratings) from the first speaker in the numerator and the summed products (of the local strength of claims and their respective relevance ratings) of all the claims in the denominator. Perhaps formulating contrasts in specific ratios to a single other claim results in output that is overly specific, and enough clash information is contained within a claim's relevance rating. In the next two experiments this prediction will be tested.

New Experiments. The following section will describe three additional experiments that test explicit predictions of the model. In all the experiments to be described in the next section, I removed subjects who had extremely strong opinions on the topics of arguments used as stimuli. In a group testing session prior to subject running (N=406 for Experiment 1, N=306 for Experiments 2 and 3), subjects indicated their prior opinion by reading statements describing the argument topics and rating their agreement with the statements on a scale from one (strongly agree) to seven (strongly disagree). An example of a statement used is “Abortion should be made illegal.” Each subject received a page featuring 16 statements, each one followed by the rating scale. I eliminated arguments and subjects for which the prior opinion means were highly biased by eliminating instances with exceptionally high means (above 5.55) or low means (below 1.50). This resulted in the elimination of two argument topics that were not used in the studies. It was necessary to set the upper bound to 5.55 instead of 5.50 in order to keep sixteen argument topics. Walton (1998) indicates that speakers who argue against the prevailing prior opinion accrue higher burden of proof than speakers who argue for prevailing opinion. In the following experiments, when fitting the burden of proof data the model can use the scores for each argument topic to account for variance due to prior belief. Appendix I shows the average prior opinion for each of the argument topics.

For each argument topic subjects also provided outcome cost ratings. A separate group of 24 subjects read the statements describing the argument topics used in the prior opinion rating task. Subjects decided the magnitude of the cost of making a decision for

each topic. In particular, we asked them to indicate “how important to society is the topic that the people are arguing about?” In the instructions, I asked subjects to consider whether the implications of invading the Middle East were different from the implications of vacuuming the carpet on Monday. Subjects rated the outcome cost on a seven point scale, with one indicating low outcome cost and seven indicating high outcome cost.

Farley (1996) argues that as the cost of being wrong in resolving the dispute gets higher, the burden of proof for the speaker forwarding an argument gets higher as well. Note that the outcome cost in informal arguments is slightly different from the formal use of the term in the judicial system. During a trial, burden of proof is designed to reflect the outcome cost of punishing the defendant. However, in informal arguments, there is no set defendant or prosecutor, and outcome cost is necessarily more general, measuring the potential implications to society more broadly. When fitting the burden of proof data the model can use the scores for each argument topic to account for variance due to outcome cost. Appendix I shows the average outcome cost for each of the argument topics.

Experiment 1 tested the fit of the model in one- and two-branch arguments, both of which were longer than those used in Pilot Experiment 2. Subjects rated local strength in isolation, the relevance of each claim in the argument, the support strength of each claim, and burden of proof. In Experiment 2, I tested the concession rules. Subjects again rated local strength, relevance, support, and burden of proof; however, they

assessed two-branch arguments either with or without concessions. In Experiment 3, I tested the effects of challenges by having subjects rate burden of proof on different versions of the same arguments, including ones with and without queries.

### Experiment 1

In this study, the model's fit is tested in one- and two-branch arguments consisting of eight claims. There were four separate groups of subjects. The first three groups provided ratings for the input to the contrast score parameters of the model. One group assessed the local strength of each claim in isolation, while the second group rated relevance scores in the argument context. The third group of subjects provided support strength ratings. Finally, the fourth group rated burden of proof in the argument context in order to provide ratings to assess the output of the model.

I predicted an anti-primacy effect, since each speaker offered the same number of claims, leaving the initiator with one less contrast than the recipient. I did not predict any burden of proof differences between one- and two-branch arguments, since the two-branch arguments did not feature any concessions and each speaker offered the same amount of contrasts. For both local strength and support strength, I predicted that the contrast scores would outperform the baselines based on a simple sum or ratio of the ratings. However, based on the results from Pilot Experiment 2, it could be the case that the local-relevance ratio (which takes into account respective local strength ratings and relevance ratings) performs as well as the local contrast model. Finally, I predicted

recency effects, in that contrasts at the end of the dispute would have the most impact on burden of proof.

### Materials

There were sixteen argument topics, each of which appeared in a one-branch and a two-branch version. The specific topics were the same ones used in the group testing session except for two that were replaced due to subjects having extremely biased opinions on them<sup>6</sup>. An example can be seen in argument 7.

- (7) a. Andy: We should raise tariffs on imported goods.
- b. Erin: Imported goods are sold throughout the country in all types of stores.
- c. Andy: Our own industries are struggling to get by because of cheap foreign products.
- d. Erin: Well we can't just go and raise tariffs whenever we feel like it.
- e. Andy: We've always raised tariffs in the past to help out our industries.
- f. Erin: The government abides by specific regulations and incremental trends when determining the tariff rate.
- g. Andy: A major determinant of the rate changes have to do with how well our products are competing--and right now foreign products are just too cheap.
- h. Erin: Well, the solution lies in our companies reducing their costs and producing cheaper products--not in us raising tariffs.

The two-branch versions of the arguments feature the second branch beginning in line f with an addendum, as can be seen in argument 8. Consequently, claims f, g, and h are different in the two versions.

- (8) a. Andy: We should raise tariffs on imported goods.
- b. Erin: Imported goods are sold throughout the country in all types of stores.
- c. Andy: Our own industries are struggling to get by because of cheap foreign products.
- d. Erin: Well we can't just go and raise tariffs whenever we feel like it.
- e. Andy: We've always raised tariffs in the past to help out our industries.
- f. Erin: Here is another reason not to raise tariffs--in the current world economy, high US tariffs may be dangerous.
- g. Andy: Tariffs are not dangerous--other countries expect us to tax imported goods.
- h. Erin: New tariffs could anger other nations and jeopardize the upcoming international trade summit.

#### Local Strength Design

After constructing the initial arguments, the claims were taken out of the argument structures and randomized with claims from other arguments. Each subject rated the list of claims in a different random order, approximately seven claims per page. Subjects rated every claim from both versions of the arguments, although the claims

common to both versions (i.e. claims one through five) were only rated once. The instructions were the same as in the pretest from Pilot Experiment 2, where the experimenter asked subjects to “read each claim and rate how convincing it is” on a seven point scale, with a one indicating “not convincing at all” and a seven indicating “very convincing”. There were 24 subjects in this condition.

#### Relevance Score Design

Each subject received a packet consisting of an instruction page and all 16 argument topics. Half the arguments were one-branch arguments and the other half were two-branch arguments. The order of the packets were arranged such that both the order of argument topics and the order of one-branch and two-branch structures was randomized. Across subjects, each topic appeared in each branch structure an equal number of times. There were 24 subjects in this condition. In the two-branch arguments subjects provided relevance ratings from claim six (the topic claim) to claim one (the reference claim). The instructions appear in Appendix II and basically described the difference between a relevant claim and a convincing claim, asking them to rate how responsive each claim was to the one before it on a seven point scale with one being not relevant at all and seven being extremely relevant (except for claim six in the two-branch arguments, which subjects compared to the first claim).

#### Burden of Proof Design

The burden of proof packets were organized exactly as were the relevance packets, with each subject seeing all sixteen topics, half one-branch arguments and half

two-branch arguments. The instructions were identical to those used in Pilot Experiments 1 and 2. There were 20 subjects in this condition.

### Support Strength Design

In the support packets, subjects rated how well each claim supported the speakers' argument as a whole. Each subject received a packet consisting of an instruction page and all 16 argument topics. Half the arguments were one-branch arguments and the other half were two-branch arguments. The arguments appeared in their full structure as depicted in argument 8. Packets were arranged in the same manner as in the other conditions. Across subjects, each topic appeared in each branch structure an equal number of times. There were 24 subjects in this condition. The specific instructions were adapted from those used by Osherson et al. (1990) and appear in Appendix III. Subjects decided how much support each claim gave the speakers on a scale from one to seven, with numbers one through three indicating support for the first speaker, numbers five through seven indicating support for the second speaker, and four indicating completely neutral claims. Ratings farther from four indicated stronger support than ratings closer to four.

### Results

Burden of proof was calculated in the same way as in Pilot Experiments 1 and 2. I multiplied a 1 by the confidence rating if the subject chose the first speaker as having more burden of proof or a -1 by the confidence rating if the subject chose the second speaker. The grand mean for burden of proof (-.40) indicated that the second speaker had

slightly higher burden, but was not significantly different from chance. There was no significant difference between the mean burden of proof for one-branch arguments ( $M = -.60$ ) and two-branch arguments ( $M = -.20$ ).

The rest of this section will be organized as follows. I calculated two separate contrast scores, one using the local strength ratings and one using the support strength ratings. For each of the two measures, burden of proof is regressed on the individual contrasts and also on the individual raw strength ratings. I predicted that the analysis including all of the contrasts would fit the burden of proof data better than the analysis including all of the raw strength ratings. In addition, for each of the two measures, I regressed burden of proof on the aggregate measures—the sum of the contrasts and the ratio of the sums of the raw strength ratings. Again, I predicted that the analysis using the contrast aggregate would outperform the analysis using the raw strength aggregate.

There were 32 data points (one- and two- branch versions for each of the sixteen separate argument topics); for each data point I averaged across subjects and computed four measures: the isolated strength of each of the eight claims, the responsiveness of the second through the eighth claims, the general support levels for the second through eighth claims, and the burden of proof score. Contrast scores were computed according to the formulas for one-branch arguments and two-branch addendum arguments described in the Model section. I first report the results for local strength contrasts.

#### Local Strength.

In this section I computed a contrast for each claim and compared the contribution of the contrasts to burden of proof to the contribution of the individual strength ratings to burden of proof. Contrasts should predict the data more effectively than the individual ratings. In a later section, I implement the third formula in the Model section and take a sum of all the contrasts.

In this first analysis, I regressed burden of proof on ten factors: the local strength ratings for the eight claims, prior opinion ratings, and outcome cost ratings. The latter two factors were included to account for variance that reflects people's average belief about the argument topics. Prior opinion was coded such that lower scores indicated agreement with the first speaker's position. Table 5 shows the standardized means of the ratings, correlations with burden of proof taken from a separate analysis from the regression, and standardized coefficients and p-values for the factors taken from the regression analysis. The coefficients in Table 5 were entered with the sign preserved. Even claims (made by the second speaker) should be positive while odd claims (made by the first speaker) should be negative.<sup>7</sup> The overall regression was significant,  $F(10, 21)=3.62$ ,  $p<.01$ ,  $R^2=.63$ ,  $RMSE=1.57$ . As Table 5 shows, none of the local strength ratings of individual claims contributed to burden of proof significantly using an alpha level of .05.

Contrasts were computed according to the formulas for one- and two-branch arguments depicted in the model section. In the next analysis, I regressed burden of proof on nine factors: the seven contrasts, prior opinion ratings, and outcome cost ratings.

Table 6 shows correlations with burden of proof, standardized coefficients, and p-values for the factors. The overall regression was significant  $F(9,22)=4.99$ ,  $p<.001$ ,  $R^2=.67$ ,  $RMSD=1.45$ . As Table 6 shows, two of the contrasts significantly predicted burden of proof (the second and the fifth), while none of the individual ratings from the previous analysis using the local strength ratings were significant. The sign of the coefficient of the fifth claim is in the wrong direction, in that a low contrast reduces the first speaker's burden of proof. This effect may result from subjects comparing the magnitude of that contrast to other (larger) contrasts made by the first speaker in the argument.

#### Support Strength.

Again, I regressed burden of proof on nine factors: the support strength ratings for the second through eighth claims, prior opinion ratings, and outcome cost ratings. Table 7 shows the standardized means, correlations with burden of proof, standardized coefficients, and p-values for the factors. The overall regression was significant,  $F(9,22)=6.87$ ,  $p<.001$ ,  $R^2=.74$ ,  $RMSD=1.29$ . As Table 7 shows, only the support strength rating for the third claim was significant using an alpha of .05. The coefficient was in the correct direction, with high support reducing burden of proof.

In the next analysis, I regressed burden of proof on nine factors: the seven contrasts, prior opinion ratings, and outcome cost ratings. Table 8 shows correlations with burden of proof, standardized coefficients, and p-values for the factors. The overall regression was significant  $F(9,22)=7.28$ ,  $p<.0001$ ,  $R^2=.75$ ,  $RMSD=1.26$ . As Table 8 shows, three of the contrasts significantly predicted burden of proof (the third, the fourth,

and the eighth), while only one of the individual ratings from the previous analysis using the support strength ratings were significant. The coefficients for the three significant contrasts were in the correct direction.

#### Comparing the Contrast Scores to Baselines.

To gauge the efficacy of the contrast model, it is necessary to compare the model to baselines that don't take into account the structure of the argument. Table 9 shows the fits of the different types of aggregate models for one- and two-branch arguments. For each type of representation, I regressed burden of proof on that representation alone (i.e., without other baselines). Consequently there is just a single parameter fitting 32 data points.

On Table 9, Local Ratio is the ratio between the sum of the local strength of speaker 1's claims and the sum of the local strength of all the claims in the argument. Support Ratio is the same measure, but with support strength ratings. Loc-Rel Ratio is the local-relevance ratio model described in Pilot Experiment 2. The Local Contrast is the aggregate contrast score computed using local strength ratings while setting all relevance scores to 1. I include this measure to demonstrate the unique impact of relevance on the contrast score. Local Contrast \* Rel is the aggregate contrast score using local strength ratings and actual relevance ratings. The Support Contrast and Support Contrast \* Rel are the same as the last two, except using support strength ratings as input instead of local strength ratings. Support SqContrast and Support SqCont\*Rel will be discussed in the General Discussion.

The data provide limited support to the contrast models. As Table 9 shows, the model using local contrasts with relevance ratings ( $\underline{R}^2 = .25$ ) accounts for more than twice the amount of variance as the Local Ratio baseline ( $\underline{R}^2 = .12$ ). However, the local-relevance ratio ( $\underline{R}^2 = .27$ ) does just as well as the contrast model. Using support ratings, the model using contrasts without relevance ratings ( $\underline{R}^2 = .58$ ) accounts for slightly more variance than any other model. It makes sense that local contrasts would benefit from relevance ratings more than support contrasts, considering that the support strength ratings were made in the argument context and somewhat take into account the relations among claims.

The aggregate contrast score using support strength ratings ( $\underline{R}^2 = .58$ ) fit the burden of proof data better than the one using the local strength ratings ( $\underline{R}^2 = .25$ ). There are a few explanations for this disparity. First, it may have been difficult for subjects to rate the local strength of claims that were presented in a random list. Many of the claims were difficult to understand once taken out of the argument structure. Consider claim c in argument 5: “You have to sit through all the side changes.” This proposition draws a large degree of its meaning from the argument context. Consequently, the local strength ratings may not have been entirely valid. This hypothesis receives some support from Figure 4, which plots local strength ratings and support strength ratings by claim number. On average, subjects rated local strength ( $\underline{M} = .60$ ) lower than support strength ( $\underline{M} = .78$ ) for the same claims,  $t(47)=3.46$ ,  $p<.05$ .

Even if the local strength rating task had been more natural, the support contrasts most likely would have outperformed the local contrasts. Since the support ratings are made within the argument context, they take into account the relation a claim has with all of the other claims in the argument, including all the claims made by a speaker's opponent. The contrasts using local strength only take into account the relationship a claim has with the specific claim it attacks. However, the support contrasts not only take into account the immediate clash relations, but also the relationship that claims have with the other claims in the argument.

Along those lines, the Support Ratio ( $\underline{R}^2 = .56$ ) did a much better job of fitting the burden of proof data than the Local Ratio ( $\underline{R}^2 = .12$ ). Since the support strength ratings measure the relations between claims while the local strength ratings do not, the simple Support Ratio fits the data well. Formulating contrasts out of the support ratings results in a fit about equal to the Support Ratio. On the other hand, the Local Contrast using relevance ratings clearly outperforms the Local Ratio. However, computing the contrast ratios is not necessary for the fit. Matching the local strength of a given claim with its respective relevance rating is sufficient, as the fit of the local-relevance ratio demonstrates.

In the analyses above, outcome cost was a significant predictor of burden of proof. The correlation was negative ( $\underline{r} = -.63$ ), such that when the first speaker began an argument with a high outcome cost, he or she had less burden of proof than when beginning an argument with a low outcome cost. To further document this effect, I

separated the arguments into two groups--high outcome cost or low outcome cost--and ran an ANOVA with outcome cost as the independent variable and average burden of proof as the dependent variable. The difference was significant,  $F(1,30)=12.86$ ,  $p<.01$ . It seems that in this experiment, subjects punished the first speaker for beginning an unimportant or trivial argument. In other words, judges are more willing to allow a speaker to begin an argument without burden when the issue being discussed has potentially large implications on the judge's life.

This effect is the opposite of what one would predict given Farley's (1996) hypothesis that as the outcome cost for the defendant increases, so does the initiator's burden of proof. However, note that these are two different types of outcome cost. Farley discusses formal arguments in which the outcome cost is measured in relation to one of the participants in the dispute. In the informal arguments used in this study, I measured outcome cost in relation to society as a whole, since it didn't make sense to measure it in relation to the arguers (who were not on trial). Consequently, the results are different from what Farley's framework would predict.

One prediction that failed was that the first speaker should have had more burden of proof overall than the second one, since that speaker offers one less contrast. With longer arguments, the impact of initiating an argument may be diminished as one of the speakers accumulates a larger contrast advantage. In other words, offering a claim without a contrast (i.e. making the first claim in an argument) may not be as damaging in

long arguments where the initiator can accrue large advantages in contrasts by the end of the dispute.

To provide evidence for this idea of argument length overwhelming the anti-primacy effect, I ran a post-test in which a separate group of subjects assessed burden of proof on arguments of varying lengths. I took the one-branch arguments from the main experiment and showed subjects either the first four claims, the first six claims, or all the claims from the arguments, with the prediction that there should be more anti-primacy in shorter disputes. There were 18 subjects in the post-test, and each subject received a packet with five arguments of each of the three lengths, resulting in 15 total arguments in the packet. Across subjects, each argument topic appeared in the three length conditions an equal number of instances. The procedure was identical to that used in the burden of proof condition from the main experiment. As predicted, there was a significant linear trend  $F(1,17)=6.02$ ,  $p<.05$ , with the highest burden of proof score for the first speaker in the four claim arguments ( $M = 1.24$ ) and the lowest burden of proof for the first speaker in the eight claim arguments ( $M = -.13$ ). Burden of proof for the six claimed disputes was between the other two conditions ( $M = 1.04$ ). In conclusion, the anti-primacy effect was strongest in the shorter arguments. In long disputes with many contrasts, the first speaker does not suffer as much from offering the initial claim without a contrast as he does in short arguments. One exception to this effect of length can be seen in Pilot Experiment 1, in which there was anti-primacy for the ten-claimed, one-branch

arguments. I would predict that there would be a great deal more anti-primacy in that study if subjects rated shortened versions of those one-branch arguments.

In summary, some of the results from Experiment 1 are in line with predictions. For both local strength and support strength, individual contrasts significantly predict the burden of proof data while individual strength ratings do not. Furthermore, for local strength, the aggregate contrast score parameter performs better than other aggregate baselines that don't take into account relevance ratings when fitting the data. However, the simple ratio model that matches local strength and relevance does just as well as the contrast model. There was some limited evidence for the recency effect, as the final support contrast was a significant predictor of burden of proof. Anti-primacy occurred, but only in short arguments. In the following experiment, I further tested the predictions of the global strength model by examining arguments structured in two branches.

## Experiment 2

In this experiment I tested the global strength model in multi-branch arguments, ones with and without concessions. As in Experiment 1, subjects rated local strength, relevance, support strength, and burden of proof for each argument.

Given the results from Experiment 1, I did not predict an anti-primacy effect for addendum arguments, since the arguments used were eight claims long. However, I did predict that in the addendum arguments the first speaker would have more burden of proof than in the concession arguments where the second speaker concedes claims and consequently accrues fewer contrasts. As in Experiment 1, I predicted that the contrasts

(both on an individual and aggregate level) would outperform the individual strength ratings and the aggregate baselines when fitting the burden of proof data. However, given the results from previous studies the local-relevance ratio should do as well as the local contrast model. In addition, in light of Experiment 1, we should expect the support contrast to outperform the local contrast model. Finally, if a recency effect is to be observed, then contrasts at the end of the argument should significantly predict burden of proof.

### Materials

There were two versions of each of the 16 two-branch argument topics. The arguments were on the same topics as in Experiment 1, but used different supporting claims to accommodate the new structures and to provide generalization for the model. In the addendum arguments, the second speaker left the first branch open when he began the second branch. Argument 8, shown above, is such an example. In the concession arguments, claim f by speaker two is changed to the following: “f. Erin: That may be true, but in the current world economy, high US tariffs may be dangerous.” Other than that change in wording, there was no difference between the two versions.

### Design/Procedure

The procedure was the same as in Experiment 1. Four separate groups of 16 subjects rated each of the four measures (one measure per group) in two blocks. Argument version (concession vs. addendum) was varied within subject. For half the subjects, the first block featured eight concession arguments and the second block

featured eight addendum arguments. For the other half of the subjects the order of blocks was reversed. Each argument topic appeared in both concession and addendum versions an equal number of times across subjects. For the relevance input measure, subjects rated how relevant claim six was to claim one.

### Results

In the main experiment, burden of proof was calculated in the same way as in previous experiments. The grand mean for burden of proof was .02 ( $SD = 5.21$ ), and was not significantly different from chance. While the difference between argument versions was not significant,  $F(1,14) = .50$ ,  $p > .10$ , as predicted the burden of proof for concessions ( $M = -.32$ ,  $SD = 5.31$ ) was on the second speaker while the burden of proof in addendums ( $M = .35$ ,  $SD = 5.14$ ) was on the first speaker.

Just as in Experiment 1, I calculated two separate contrast scores, one using the local strength ratings and one using the support strength ratings. Again, for each measure I ran analyses using the individual measures and the aggregate measures. Since there were two versions (concession and addendum) of each of the sixteen separate argument topics, there were 32 data points to use while testing the model. For each data point I averaged across subjects and computed four measures: the local strength of each of the eight claims, the relevance of the second through the eighth claims, the support strength of the second through eighth claims, and the burden of proof score. Contrast scores were computed according to the formulas for two-branch arguments described in the Model section. I first report the results for local strength contrasts.

Local Strength. In the first analysis, I regressed burden of proof on ten factors: the local strength ratings for the eight claims, prior opinion ratings, and outcome cost ratings. Table 10 shows the standardized means, correlations with burden of proof, standardized coefficients, and p-values for the factors. The overall regression was significant,  $F(10,21)=3.87$ ,  $p<.01$ ,  $R^2=.65$ ,  $RMSD=1.78$ . As Table 10 shows, none of the local strength ratings of individual claims contributed to burden of proof significantly.

Contrasts were computed according to the formulas for concession and addendum arguments depicted in the model section. For concession arguments, the second speaker did not gain a contrast for claim d (I will discuss this exclusion in more detail later in this section). In the next analysis, I regressed burden of proof on nine factors: the seven contrasts, prior opinion ratings, and outcome cost ratings. Table 11 shows correlations with burden of proof, standardized coefficients, and p-values for the factors. The overall regression was significant,  $F(9,22)=4.48$ ,  $p<.001$ ,  $R^2=.65$ ,  $RMSD=1.74$ . As Table 11 shows, two of the contrasts significantly predicted burden of proof (the third and the fifth), while none of the individual ratings from the previous analysis using the local strength ratings were significant. The significant contrasts were in the correct direction. Furthermore, even though this analysis using the contrasts had one less parameter than the analysis using the local strength ratings, the  $R^2$  values were equal.

Support Strength. To analyze support strength, I regressed burden of proof on nine factors: the support strength ratings for the second through eighth claims, prior opinion ratings, and outcome cost ratings. Table 12 shows the standardized means,

correlations with burden of proof, standardized coefficients, and p-values for the factors. The overall regression was significant,  $F(9,22)=4.62$ ,  $p<.001$ ,  $R^2=.65$ ,  $\text{RMSE}=1.73$ . As Table 12 shows, none of the support strength ratings were significant using an alpha of .05.

In the next analysis, I regressed burden of proof on nine factors: the seven contrasts, prior opinion ratings, and outcome ratings. Table 13 shows correlations with burden of proof, standardized coefficients, and p-values for the factors. The overall regression was significant  $F(9,22)=5.54$ ,  $p<.001$ ,  $R^2=.69$ ,  $\text{RMSE}=1.63$ . As Table 13 shows, the second contrast predicted burden of proof, while none of the individual ratings from the previous analysis using the support strength ratings were significant. However, that contrast is in the wrong direction, in that a high contrast increases the second speaker's burden of proof. This result is surprising, and may have occurred due to subjects having difficulty rating the support of the second claim in the argument (since the second speaker does not have an argument yet to support). On average, however, both speakers' coefficients for contrasts were in the predicted direction in Experiments 1 and 2.

Comparing the Contrast Scores to Baselines. I compared the contrast model to the same baselines used in Experiment 1. Table 9 shows the fits of the different types of models for concession and addendum arguments. As in the previous experiment, the data from Experiment 2 provide a limited support the contrast model. Table 9 shows that the model using local contrasts with relevance ratings ( $R^2 = .28$ ) accounts for almost three

times the amount of variance as the Local Ratio baseline ( $\underline{R}^2 = .10$ ). Again, however, the local-relevance ratio fits just as well as the local contrast model ( $\underline{R}^2 = .28$ ), demonstrating that contrasts are not necessary when relevance and local strength are matched.

Using support ratings, the contrast model using the relevance ratings accounts for more variance than any other model ( $\underline{R}^2 = .38$ ). As in the previous experiment, the contrasts using support strength ratings did quite a better job fitting the burden of proof data than the contrasts using the local strength ratings. Furthermore, as in Experiment 1, the Support Ratio baseline did a much better job of fitting the burden of proof data than the Local Ratio baseline, since subjects rated support strength in the argument context. This is an issue that will be discussed in detail in the General Discussion. Figure 5 plots local strength ratings and support strength ratings by claim number. As in Experiment 1, subjects rated local strength ( $\underline{M} = .56$ ) lower than support strength ( $\underline{M} = .76$ ) for the same claims,  $t(31)=4.02$ ,  $p<.05$

One of the most striking patterns in Table 9 is the difference between concession arguments and addendum arguments. Averaged across all the different types of models,  $\underline{R}^2$  for concession arguments is .44, while  $\underline{R}^2$  for addendum arguments is .11. This result is surprising, especially considering that there is not a significant difference in burden of proof for the two types of arguments in this study. Furthermore, the addendum arguments from this study are similar to the two-branch arguments in Experiment 1. Both studies used the same argument topics, but I changed the supporting claims made by

each speaker for the arguments in Experiment 2. For some reason those changes made the models fit the burden of proof data poorly.

The disparity was not due to a difference in burden of proof assignment--the ratings from the two studies correlated significantly by topic,  $r=.69$ ,  $p<.05$ .

Consequently, the models must have fit poorly due to the individual ratings. There were no large differences in local strength or support strength ratings by position, however, the correlation between relevance ratings of the two experiments by position was low  $r=.26$ . Perhaps this disparity is due to subjects having difficulty formulating relevance ratings in the second study; the changes I made in the supporting claims may have resulted in a more difficult relevance judgment task.

There was a non-significant trend for anti-primacy in addendum structures but not for concessions. Previous research (Rips, 1998) demonstrates that subjects don't always adhere to rebutting principle (a). As previously noted, in argument 1, Billy concedes that claim e is true by accepting a defeater. Rebutting principle (a) dictates that Billy rejects his own claim d. The current model is designed such that Billy does not get a contrast for that claim. However, in Rips's studies, subjects on average did not rate Billy as uncommitted to claims similar to claim d (although the mutually rejected claim was rated as 'less committed' than other claims). To test this principle further, I fit the contrast model two different ways: including the contrast for claim d and not including the contrast for claim d. The difference between the two fits was small, but in the direction that supports rebutting principle (a). When the contrast for d is included,  $R^2$  for the

support contrast drops from .38 to .36. Consequently, the non-significant trend of anti-primacy in addendum (but not concession) arguments, coupled with the poorer fit for the model that included the mutually rejected claim provide support for Rebutting Principle (a).

The prediction that later contrasts in the argument should have the highest impact on burden of proof was not supported by the data in this study. Apparently the recency effect is not as robust a phenomenon as the other argument characteristics. In general, results from Experiment 2 provide limited support for the contrast model. Individual contrasts significantly predicted burden of proof while the individual strength ratings usually did not. Moreover, for both local strength and support strength, the aggregate measures based on contrasts fit the burden of proof data better than the simple ratio aggregates that did not include relevance ratings. As in previous studies, however, the local-relevance model fit the data just as well as the local contrast model. Therefore, computing contrast ratios may not be necessary in determining global strength. In the following experiment, I tested the effects of query on burden of proof.

### Experiment 3

In this study I isolated the effects of challenges in conversational arguments. Subjects read short arguments between two speakers. Appendix IV shows the content of the arguments. Each argument appeared in two structures, one where the second speaker makes a challenge to the first claim (challenge) and one where the second speaker offers a direct rebuttal to the first claim (direct rebuttal). The two structures are represented in

Figure 2. Note that even though the speakers offer the same number of substantive claims in both conditions, in the challenge structure the speaker who makes the query changes the course of the dialogue such that the other speaker gains one less contrast. In the challenge condition there is only one contrast at claim B. In the direct rebuttal, each speaker gets a contrast, one at claim B and one at claim A<sub>1</sub>. In the figure, bold arrows represent contrasts.

I also manipulated contrast direction such that the contrast (between claim A<sub>1</sub> and claim B in Figure 2) could either favor the first speaker or the second speaker. In Appendix IV, the argument between Sean and Kary about communication in relationships is an example of the contrast favoring the first speaker. The argument between Tom and Ron about the job market is an example of the contrast favoring the second speaker.

Initially, I predicted a main effect for contrast direction, in that the speaker with the contrast advantage should have less burden of proof. In addition, I predicted a main effect of structure, in that the speaker who offers B should have lower burden of proof in the challenge condition since he has one more contrast than his opponent in this condition.

### Materials

The materials were 16 short argument topics consisting of three claims each. There were two within subject variables. The first was argument structure, either challenge or direct rebuttal. The second variable was the direction of the contrast, either

in for-B or against-B. A pretest ensured that in the for-B condition, the local strength of B was greater than the local strength of A<sub>1</sub>, the claim with which it clashes. Similarly, in the against-B condition, the local strength of B was lower than the local strength of A<sub>1</sub>.

To determine contrast direction, I pretested a separate group of nine subjects who rated the local strength of all the claims in isolation on a seven-point scale. The average differences between local claim strength were in the correct direction for the stimuli. Claim A<sub>1</sub> was on average 1.68 higher than claim B in the against-B condition (SD = .54) and claim B was on average 1.62 higher than claim A<sub>1</sub> advantage in the for-B condition (SD = .51). A paired t-test indicated that there was no difference in the magnitudes of the advantages for the two conditions,  $t(8)=.26$ ,  $p>.05$ , and all differences between claims for all topics in both conditions were in the predicted direction and were greater than 1.0. Appendix IV shows the average local strength scores for each claim.

#### Design and Procedure

Subjects read the short arguments and rated burden of proof. The experimenter instructed subjects that they only saw a portion of the argument; the beginnings and ends of the section did not necessarily reflect the beginnings and ends of the argument. In this fashion I hoped to mitigate the effects of having opposite speakers presenting the final claim in the dispute. In other words, in the challenge version, speaker Y gets the final speaking turn, and in the direct rebuttal version, speaker X gets the final speaking turn. However, previous research has shown that neither offering an extra non-claim speaking turn nor the final claim affects burden of proof in disputes (Bailenson & Rips, 1996).

Regardless, I instructed subjects that the last claim on the page was not necessarily the final speaking turn in the dispute.

Subjects saw all 16 argument topics, which are listed in Appendix IV. In each packet, eight of the topics were for-B and eight were against-B. Subjects assessed four of each contrast type in each of the two structures. Order of both contrast condition and argument structure was randomized, and across subjects each argument appeared in each structure an equal number of times. When subjects had completed the packets the experimenter asked them to go back to each argument and to describe the reasons why they chose the selected speaker as having more burden of proof.

### Subjects

The subjects were 28 Northwestern University undergraduate students who participated in order to gain partial credit in an introductory psychology course.

### Results and Discussion

The dependent variable was the percentage of subjects who chose the first speaker as having more burden of proof. As predicted, there was a main effect for contrast direction; the speaker with the contrast advantage was chosen as having burden of proof less often (28.93%) than the speaker without the advantage (71.07%),  $F(1, 27)=95.52$ ,  $p<.001$ . Furthermore, as predicted there was a main effect for structure in that the second speaker was chosen less often as having the burden of proof in the challenge condition (43.64%) than in the direct rebuttal condition (57.52%),  $F(1,27)=5.22$ ,  $p<.05$ . The interaction was not significant.

Results from the justification task for burden of proof can be seen in Table 14. Two judges classified the list of justifications and reliability in terms of agreement was 92%. The two judges discussed items on which they differed and tried to reach consensus. Those items that remained under contention were discarded.

Justifications were placed into one of six categories: anti-primacy (i.e., ‘John spoke first so he has to prove his position’), claim strength (i.e., ‘Ted had stronger arguments’), support (i.e., ‘Bob presented more ideas to support himself’), personal agreement (i.e., ‘I just think that she is right’), evidence (i.e., ‘Alan doesn’t show any evidence for his argument’) and relevance (i.e., ‘Vince’s arguments are irrelevant’). The results indicate that subjects explicitly attend to claim strength while making burden of proof judgments. While the majority of the responses fell into agreement and support, subjects also cited relevance, and even anti-primacy as reasons for their decisions.

The results of Experiment 3 provide a great deal of support for the contrast model. In arguments where the contrast advantage favored a given speaker, subjects chose that speaker as having less burden of proof than his or her opponent. While this effect could also be motivated by total claim strength for each side in the argument, the two argument structures distinguish between total claim strength and contrast. In arguments where the second speaker used the query to gain an extra contrast (challenge condition), the first speaker had high burden of proof.

## GENERAL DISCUSSION

In this paper I present a global strength model that accounts for how people assess the aggregate strength of a speaker's position in conversational disputes. In arguments, speakers attempt to reduce their burden of proof by building a strong position. The model relies on the notion of contrast score, a measure that takes into account both the structure of the argument and the strength of the component claims in order to determine the overall strength of each speaker's position. Claims can be measured by their local strength (the intrinsic value of the explanation and evidence within the claim) or by their support strength (the amount of support the claim offers to the other claims in the dispute). Support strength takes into account the local strength of a claim as well as the relationship between that claim and the argument as a whole.

Contrast score is determined by three major factors: the convincingness of the claims in the argument, the ways in which the claims relate to one another, and the levels of commitment in the dispute. Contrasts are computed by comparing each proposition to the proposition with which it clashes. Conversational moves such as queries and concessions affect commitment levels and clash relations, consequently impacting contrast scores and global strength. Contrasts can be computed using local strength or support strength.

The paper presents five experiments that provide support for the model. The support can be seen in two general forms. First, the model predicts both intuitive and non-intuitive burden of proof phenomena that occur in conversational disputes. Second, the model significantly fits the burden of proof data. However, in some instances, other representations perform as well as do the contrast scores. In this section I will summarize the main results from the experiment and then will discuss reasons why the contrast model did not always outperform similar models that do not take argument structure into account. I will also explain why none of the models can account for more than two-thirds of the variance.

### Predicting Phenomena

Assessment of burden of proof in conversational arguments depends on certain characteristics of the argument. First of all, the initiating speaker who begins the argument makes a claim that does not clash with other claims. As a result, in arguments where speakers make the same number of claims, the initiator's global strength is lower and burden of proof is higher since he offers one less contrast than the recipient (the second speaker). Evidence for this anti-primacy effect can be seen in Pilot Experiment 1, Pilot Experiment 2, and in Experiment 1. In addition, there is ample evidence from previous studies which document this phenomenon (Rips, 1998; Bailenson, 1997; Bailenson & Rips, 1996) as well as other models that have incorporated the phenomenon into their architecture (McConachy, Korb & Zukerman, 1998).

However, even in arguments where the two speakers present an equal number of claims, anti-primacy does not always occur. In certain situations, the model predicts that the burden of proof should shift to the second speaker. One instance where this shift occurs is when the second speaker concedes claims. By retracting one of his claims, the recipient propels that claim into a state of mutual rejection, and loses the contrast from the mutually rejected claim. Evidence for this shift in burden of proof can be seen in Pilot Experiment 1 and in Experiment 2. Another way to accrue more contrasts is to use conversational moves such as queries to remove an opportunity for your opponent to offer a contrasting claim (Experiment 3).

The anti-primacy effect can also be mediated by the perceived impact of the outcome of the dispute. Data from Experiment 1 demonstrate that in arguments where judges view the outcome of an argument to be extremely important, the first speaker does not necessarily have more burden of proof than the second speaker, even if he tallies one less contrast. If the need to resolve an argument seems dire to the judge, then the speaker that initiates the dispute is not necessarily punished with higher burden. In addition, the anti-primacy effect can be overwhelmed a large number of contrasts that contribute to the overall score (Experiment 1, post-test), since in longer arguments the first speaker has more of an opportunity to make up for the contrast he missed by making the first claim. This is consistent with results from Bailenson & Rips (1996), who found anti-primacy in five-claimed arguments.

Another reason that the first speaker may have more burden of proof than other speakers has to do with the information contained in the first claim of the argument. Often times the first claim of an argument is a simple statement of position that does not contain any actual explanations or evidence for the statement. Consider the arguments in Appendix IV, which have the exact same first claim as the arguments from Experiments 1 and 2. In only three of those arguments (topics 3, 9, and 11) does the first speaker provide elaboration in the first claim. For example, in topic 11, Elaine argues that ‘Capitol punishment should be televised in order to prevent future crimes. This is different from the other 13 arguments in which the first speaker concisely states his position without providing reason.

In order to test the effect of providing reason in the first claim, I combined the subjects from Experiments 1 and 2 and computed two means for each subject—one for the three arguments that contained qualification and one for the other arguments. A paired t-test indicated the first speaker had less burden of proof for arguments in which his first claim contained qualification ( $M = -.87$ ) than in arguments in which his first claim did not contain qualification ( $M = -.13$ ),  $t(35)=2.85$ ,  $p<.05$ .

In the current work, as well as in previous studies, most arguments used have featured first claims without explanations. For example, Rips (1998) found an anti-primacy effect even though the arguments used in his studies were relatively long (nine claims). However, the first speaker presented qualification in the first claim in only one of the twelve arguments from those studies. Likewise, all the arguments used in

Bailenson & Rips (1996) contained arguments without explanations in the first claim. In future work, the difference between presenting reason in the first claim and not presenting reason needs to be systematically explored.

### Comparison to Alternatives

In addition to predicting phenomena that occur in conversational arguments, I formally fit the model and compared that fit to alternative representations. One class of alternatives discussed in previous sections is representations that take the ratio of the strength of each speaker's claims without taking into account the clash relations between the claims. In many instances, for both local strength and support strength, the contrast score parameter marginally outperformed the simple ratio baseline (Pilot Experiment 2, Experiment 1, Experiment 2). Because the contrast scores take into account the structure of the argument and the relations among the claims, they accounted for slightly more variance than the aggregates that simply sum strength ratings. However, when the ratio baselines are modified to include relevance ratings (i.e., local-relevance ratio) they perform just as well as the contrast model. Consequently computing local ratios may not be necessary, seeing as the relevance ratings measure the same amount of clash between two claims as do the formal contrast comparisons. In future work I would gather pairwise relevance ratings between all the claims in the argument (as opposed to only the previous claim). When building the Global Strength Model, I would then incorporate less structure into the model (i.e., not include contrast ratio comparisons). Instead I

would merely pair the multiple relevance ratings with the local strength of the claims, as in the local-relevance ratio.

One different way to compute contrasts is by squaring all of the ratings before taking a ratio. In this sense, the differences between clashing claims are magnified and contrasts are pulled away from one half. Table 9 shows the results of an analysis that regressed this type of model on burden of proof. SupportSq Contrast is a contrast score that uses support ratings, but squares them before taking a ratio. For this measure, all relevance scores are set at 1. Support SqCont \*Rel is the same measure as above, but using the actual relevance ratings. As Table 9 shows, this representation fits better than the simple ratio baselines but not as well as the original contrasts. In the current studies, there was a high degree of variance in claim strength. Perhaps in arguments where opponents made claims of highly comparable strength, this representation would be more useful.

Another alternative representation that I explored was a slight modification of Kahneman's model of peak/final dominance in evaluations of aversive experiences (Kahneman, Fredrickson, Schreiber & Redelmeier, 1993; Redelmeier & Kahneman, 1996). According to this approach, when people retrospectively evaluate an aversive experience, they predominantly attend to the peak of their aversion and to the final stage of their aversion. Their studies present convincing evidence that these two factors are used substantially more during evaluation than other factors, such as the duration of the experience or the total amount of aversion.

Even though the peak/final dominance model was designed to address processes vastly distinct from burden of proof assessment, the model makes sense when modified and applied to conversational arguments. I examined the data from Experiment 1 and Experiment 2 and tested a modification of the Kahneman model. In the modification, burden of proof was regressed on only two factors for each argument topic—the peak contrast (the highest contrast for each argument topic) and the final contrast. Across strength measures, the fit of the Kahneman model was lower than the fit of the contrast models in both Experiment 1 ( $R^2=.13$ ) and Experiment 2 ( $R^2=.15$ ). Neither of the fits was significant.

One reason for this poor fit may be the manner in which subjects assessed the arguments. In Kahneman's aversion studies, subjects assessed the amount of aversion retrospectively. Consequently, the influence of the peak contrast may be strongest when judges are assessing arguments from memory. This discrepancy between judging arguments as they are being encoded and judging arguments retrospectively has been discussed in previous work by Hogarth and Einhorn (1992). In future studies, the peak contrast may have a larger impact if subjects delay assessment for some period of time after the arguments have finished.

In both Experiment 1 and Experiment 2, the difference in fit between support contrasts and support ratings was not as large as the difference between local contrasts and local ratings. Since support ratings take into account the relationship between a given claim and the other claims in the argument, they already measure some degree of

argument structure. Consequently, computing contrasts only marginally improves the fit of support ratings. In the next section, I will explore at length the difference between the two types of ratings, and will describe a pilot study that provides some insight as to why the support models outperformed the local models.

Comparing Local Strength and Support Strength.

In both Experiment 1 and Experiment 2, the models that utilized support strength fit the data better than models that utilized local strength. Earlier, I discussed reasons for this disparity. One reason may be due to the difficulty subjects had when rating the local strength of claims out of the argument context. This idea is supported in Figure 4 and Figure 5, in which local strength ratings overall are rated lower than support strength ratings. Furthermore, one would expect support strength to be more informative than local strength, since subjects take into account local strength when they rate support strength.

Another reason support ratings were more successful than other ratings may be due to the fact that support strength ratings take into account relations between a given claim and all the opposing claims in the argument. Consider Erin's strategy in argument 9:

- (9) a. Andy: We should raise tariffs on imported goods.
- b. Erin: No we shouldn't—tariffs are too high already.
- c. Andy: But our own industries are struggling to get by because of cheap foreign products.

d. Erin: Row boats shouldn't go out in storms.

e. Andy: We've always raised tariffs in the past to help out our industries.

In this discussion, Erin avoids clash by making a claim that has nothing to do with raising tariffs. In claim e, Andy makes a choice. Instead of responding to claim d with an undercutting defeater (i.e., 'Rowboats have nothing to do with tariffs. '), he continues to develop the line of reasoning he began in line a and chooses to clash with Erin's claim b. In this situation, the local contrast score does not effectively reflect Andy's burden of proof advantage because Andy's score might suffer due to a low relevance score between claim e and claim d. Furthermore, the contrast model does not measure Andy's clash relations between claim e and claim b. On the other hand, the support ratings made within the argument context become especially useful because they partly take into account the relationship between a claim and all the opposing claims that came before it. Consequently, it is more equipped to model an anomalous argument such as argument 9.

Similarly, it could be possible that the models based on support ratings fit the burden of proof data better than other models simply because they were made in the context of opposing claims, and not because support is a more diagnostic measure of burden of proof. Because subjects were seeing the argument in its entirety, as they rated the claims they may have been computing a 'running total'. In other words, ratings near the end of the argument may have contained information about all of the opposing and supporting claims made earlier in the argument.

In Appendix V, I describe a pilot study in which subjects provided one-sided support ratings. For each side of the arguments used in Experiment 1, subjects read a sequenced monologue depicting one side of an argument. In other words, subjects rated only the claims from a single speaker, without ever seeing the opposing claims from that argument. Consequently it becomes possible to distinguish between ratings that measure support for one's own claims and ratings that measure attacks towards opposing claims in the argument structure.

Using these one-sided ratings to fit the contrast model, I attempted to demonstrate a more unique contribution of support. If the contrast model using the one-sided support ratings works as well as the model using the contextual support ratings from Experiment 1, then support strength is a more accurate measure of burden of proof than local strength. If the model using the one-sided ratings does not work as well as the model using the contextual ratings, then the support models in Table 9 are superior in part due to the presence of opposing claims within the structure, and subjects may have been computing a running strength total as they proceeded along the dispute.

Appendix V shows that contrasts computed with one-sided support ratings fit the burden of proof data no better than contrasts computed with local strength ratings. Consequently, the contextual support models listed in Table 9 may fit the data better than other models due to the presence of opposing claims. Originally, the purpose of designing the contrast model to take into account local strength and relevance (to a single clashing claim) was to isolate the processes people utilize when assigning burden of

proof in arguments. However, there is a fundamental problem with such a formulation, in that claims don't only clash with a single claim. This problem is clearly highlighted by the poor fits of the local strength contrast models. The data from these studies show that in a conversational dispute, claims actually clash with many opposing claims within the argument structure. As a result, the contrast model that takes into account only relevance to a single claim is at a disadvantage.

On the other hand, the support ratings made in the argument context can indirectly measure a claim's relations with all opposing claims, and accordingly measure burden of proof more successfully. The problem, however, with contextual support ratings is the difficulty in measuring exactly what subjects are doing when rating claims in the overall structure. Later in this section I describe a potential class of studies that would provide as much information as the contextual ratings but still allow the experimenter to monitor subjects' actions.

The local contrast score (using only relevance ratings to a single claim) might be most central to burden of proof when there is a large focus on clash in the argument. For example, in televised debates and trial settings (where there is an explicit adversarial and back environment) judges may rely more on immediate clash, and may be less concerned with comparing a claim to opposing claims that occurred much earlier on in the dispute.

Moreover, it is not always possible to get a measurement of claims within an argument structure. The contrast using local strength is especially useful for fitting arguments in situations where claims have not previously been configured in a

conversation. In other words, if one is trying to predict the burden of proof for a conversational argument that has not yet occurred, then the only measurement of strength can be the local strength (since the claims have never appeared in an overall relational structure). Consequently, local strength can be particularly useful for predicting burden of proof in new arguments as well as for designing arguments with the goal of minimizing burden of proof.

#### Why Can't any of the Models Explain all the Burden of Proof Data?

The models from the current study significantly explained the variance in the burden of proof data. However, there is still a great deal of variance that is not explained by any of the models that appear in Table 9. To explore this leftover variance, I analyzed the residual data from both Experiment 1 and Experiment 2.

Each subject gave a specific burden of proof rating for each argument topic. It could be the case that all the subjects varied systematically (above what is explained by the models) on particular argument topics. To test this, I took the difference between a subject's normalized burden of proof rating for each topic and the normalized rating predicted by the model for that topic. Consequently, for each subject there were 16 residual data points—the difference between actual ratings and ratings predicted by the model for each topic.

Principal components analysis was then performed on the inter-argument correlation matrices to determine whether or not there was a single driving factor explaining the residual variance. A single factor solution results if a single dimension

underlies patterns of agreement within a domain. Agreement occurs if: (1) the first eigenvalue is notably (three times) larger than the second and accounts for much of the variance, and (2) the first factor eigenvalue loadings (unrotated factor scores) are all positive (Romney, Batchelder, & Weller, 1986; Medin et al. 1997). If there is a single factor solution for the residual data, then the models are missing some fundamental and unified aspect of the burden of proof data. If there is not a solution, then the model is having difficulty due to subjects' widely variant strategies for determining burden of proof.

For each experiment, I used the model that accounted for the most variance (Support Ratio in Experiment 1 and Support Contrast in Experiment 2) to obtain the residual burden of proof data. For each experiment I compare the first eigenvalue to the second, I check to see if all eigenvalues for the first factor are positive, and I attempt to explain the variance of that factor.

In Experiment 1, the first eigenvalue (3.18) accounted for 19.90 % of the variance and was not three times (significantly) larger than the second (2.53), which accounted for 15.90 % of the variance. Moreover, only half of the eigenvectors were positive. In order to determine what might be driving the first factor, I ran a correlation between prior opinion ratings and the first factor eigenvectors. The analysis was significant,  $r=.71$ ,  $p<.01$ . The correlation between the remaining factors and outcome cost ratings was not significant. Furthermore, gender of the first speaker was not correlated with any of the factors. In sum, there was not a single factor solution for the residual data in Experiment

1, and the small amount of variance explained by the first factor seems to be subjects' prior opinion.

In Experiment 2, the first eigenvalue (5.13) accounted for 32.10 % of the variance and was not three times larger than the second (3.53), which accounted for 22.10 % of the variance. Moreover, only eleven of the eigenvectors were positive. A correlation between prior opinion ratings and the first factor eigenvectors again was significant,  $r=.77$ ,  $p<.001$ . The correlation between the remaining factors and outcome cost ratings was not significant and gender of the first speaker was not correlated with any of the factors. In sum, there was not a single factor solution for the residual data in Experiment 2, although more of the variance was explained by the prior opinion factor in the second study. This is consistent with the previous regression results that show: 1) lower fits for the model in the second study and 2) prior opinion to be a significant factor in Experiment 2 (as shown in Table 10) but not in Experiment 1.

In conclusion, the models in Table 9 did not account for more than two thirds of the variance in the burden of proof data. However, the variance not explained by the model cannot be significantly explained by any one factor, although some residual variance can be explained by prior opinion.

Subjects rely on many factors when determining burden of proof. In the introduction, one of these factors I discussed was the social characteristics of the arguers. It could be the case that subjects were reading into the characters in the dispute, and sometimes relying on largely orthogonal social strategies when assigning burden of

proof. Moreover, the relatively low fits may have to do with the differences between groups of subjects. Separate groups rated burden of proof and claim strength. Perhaps the model would fit more effectively if the same subjects rated burden of proof as well as the inputs to the contrast scores.

#### Explaining Non-Complementary Judgments

Previous research on burden of proof has shown that subjects do not resolve arguments in a complementary fashion (Bailenson & Rips, 1996; Van Wallendael, 1990). In other words if there is a certain argument structure that favors one speaker, that structure does not necessarily count against the other speaker to the same degree.

The current model can explain this counter-intuitive finding. The manner in which the contrast score is determined allows for the computation of a specific value for each speaker, even though the model takes the difference of those two values to output a single score. Each speaker has a certain contrast score and it is not the case the two values are necessarily of the same magnitude (but only of a different valence). It follows from that assumption that arguments are not necessarily zero-sum. For example, consider a situation where both speakers offer extremely unconvincing claims. One speaker's claims may be less convincing than the other's, but according to the model the other speaker doesn't profit only because the losing speaker's claims are worse. Instead, each speaker receives an independent global strength score based on the independent input parameters.

#### Avenues for further research

As described earlier, computing contrasts only minimally improves the fit over other types of models. Information about relevance and strength ratings do as good a job as contrast models when fitting burden of proof data. Furthermore, ratings made in the argument context approximate burden of proof ratings better than other ratings. Unfortunately, those ratings tell us little about what people are actually doing when computing burden of proof. In future studies I would elicit pairwise relevance relations between all of the claims in the dispute, and combine them with strength ratings in a general manner (i.e., the local-relevance ratio). In this manner, all information about opposing claims is available and it is possible to monitor subjects assessments of relevance. While I would keep structural information about the conversation within the model such as concessions and queries, I would not compute specific contrast ratios, since the current experiments show that relevance ratings contain sufficient amounts of clash information.

Another improvement would be to generalize the model's success by predicting the outcomes of arguments that actually occurred in the real world. Other researchers (Rips, 1998; Thagard, 1989) have successfully applied their models to arguments occurring in natural discourse. One particularly intriguing application would be to use the model to compare the arguments used in the O.J. Simpson criminal trial to the arguments used in the O.J. Simpson civil trial, seeing as the verdict ultimately fell against the defendant in one but not the other. Subjects could provide ratings for the support strength, local strength and for the relevance scores for each major proposition in the

trial. Furthermore, it may be possible to access historical data to somehow gauge the prior opinion of the juries in order to filter the variance of prevailing attitude towards the prosecution.

One crucial factor in explaining the differential outcomes of the disputes should be the outcome cost, since in one trial the defendant faced a lifetime in jail while in the other trial there were only financial implications. Generally speaking, the first claim of the trial argument made by the initiator (the prosecutor) was something along the lines of ‘Simpson is guilty of killing his ex-wife’. The anti-primacy effect resulting from beginning the dispute was more substantial in the criminal trial since, presumably, the outcome cost (for the defendant) of going to prison is higher than the outcome cost of losing money. Outcome cost modulates the anti-primacy effect (Experiment 1), and the resultant shift in the civil trial might have been enough to produce a guilty verdict.

The prospect of testing the model on an argument as extensive as a criminal proceeding raises additional issues. Certainly the criminal trial is characterized by scores of independent sub-arguments and also by hundreds of claims offered by multiple agents. Thus far the model has been used exclusively to examine disputes between two speakers that end after less than a dozen claims. However, the model is equipped with rules that can accommodate these larger and more complex structures, and it would be a worthy challenge to apply it to more substantial disputes.

### Conclusions and Implications

During a conversation, speakers who disagree about a certain issue often attempt to convince each other that their opinion reflects the truth. These types of arguments housed in dialogues are both similar and different from the one-sided arguments (composed of premises supporting conclusions) that have been predominantly studied in psychology.

The two types of arguments share certain characteristics. For one thing, both can be characterized as an attempt to persuade a judge or to determine the truth. Furthermore, both types of arguments are at least partly governed by support relations between propositions. The one sided arguments featuring premises and conclusions have been characterized as being almost exclusively dictated by support (Toulmin, 1958; Osherson et al., 1990). The current studies demonstrate that when judges assess conversational arguments, they do in fact attend to the support connections between claims.

However, support relations are only a single factor that judges of conversational disputes consider. Arguments occurring in a dialogue context are unique because their structure reflects not only the evidence and explanations within the propositions, but also the structural features of the discourse. In a conversational dispute, speakers use language to forge relations between claims (Reichman-Adar, 1984) and also to indicate their commitment to the propositions in the dispute (Rips, 1998). Moreover, in disputes between multiple agents, burden of proof can shift from one speaker to the other

depending on a variety of conversational moves. The current model explores these conversational moves and sheds light on how people reason in everyday situations.

The results from the current experiments present challenges for previous models of argument processing. For a model to thoroughly address burden of proof decisions, it should account for the types of phenomena that occur when an audience judges a conversational argument, such as anti-primacy, non-complementary judgments, the effects of conversational moves such as queries and concessions, and the importance of clash. In the beginning of this paper, I discussed how certain models could be modified to address the uniqueness of conversational disputes. The global strength model attempts to isolate the factors that contribute to burden of proof decisions. While the models did not always account for large amounts of variance in the data, they did provide some insights as to the types of factors people rely on when determining burden of proof. Existing models could be improved by adapting a similar approach.

## TABLES

Table 1. Mean contextual strength ratings in Pilot Experiment 2 by position for speaker one and speaker two.

---

Position	Speaker	
	1	2
Initial Claims	2.67	2.58
Middle Claims	4.49	4.47
Final Claims	4.73	4.83

---

Table 2. Normalized Means, correlations with burden of proof, standardized coefficients, and p-values for local strength ratings that are regressed on burden of proof for Pilot Experiment 2. The overall regression was not significant,  $F(6,2)=3.13$ ,  $R^2=.90$ , RMSD=1.55.

Claim	Mean Rating	Correlation	Coefficients	P-values
1	.46	.05	.59	.32
2	.41	.29	-.67	.25
3	.56	-.33	-.63	.29
4	.60	.17	.83	.18
5	.59	-.78	-.88	.10
6	.66	.07	.39	.36

Table 3. Normalized Means, correlations with burden of proof, standardized coefficients, and p-values for relevance ratings that are regressed on burden of proof for Pilot Experiment 2. The overall regression was not significant,  $F(5,3)=.52$ ,  $R^2=.46$ , RMSD=3.00.

Claim	Mean Rating	Correlation	Coefficients	P-values
2	.39	.45	-.07	.95
3	.67	-.05	-.06	.92
4	.70	-.03	-.15	.76
5	.71	-.61	-.66	.53
6	.70	.24	.30	.58

Table 4. Correlations with burden of proof, standardized coefficients, and p-values for contrasts that are regressed on burden of proof for Pilot Experiment 2. The overall regression was significant,  $F(5,3)=8.69$ ,  $p<.05$ ,  $R^2=.94$ , RMSE=1.04.

---

Contrast	Correlation	Coefficients	P-values
2	.51	-.04	.92
3	-.35	.04	.92
4	.38	.48	.20
5	-.77	-.35	.29
6	.73	.71	.05

---

Table 5. Normalized Means, correlations with burden of proof, standardized coefficients, and p-values for local strength ratings, prior opinion ratings, and outcome cost ratings that are regressed on burden of proof for Experiment 1. The overall regression was significant,  $F(10, 21)=3.62$ ,  $p<.01$ ,  $R^2=.63$ ,  $\text{RMSE}=1.57$ .

Claim	Mean Rating	Correlation	Coefficients	P-values
1	.51	-.29	-.01	.97
2	.54	.30	.35	.24
3	.65	.11	.01	.95
4	.60	-.25	-.38	.08
5	.62	-.60	-.29	.15
6	.63	.32	.03	.84
7	.64	-.20	.06	.72
8	.64	.04	-.12	.44
Prior Opinion	N/A	.25	.21	.32
Outcome Cost	N/A	-.63	-.33	.08

Table 6. Correlations with burden of proof, standardized coefficients, and p-values for local contrasts, prior opinion ratings, and outcome cost ratings that are regressed on burden of proof for Experiment 1. The overall regression was significant,  $F(9,22)=4.99$ ,  $p<.001$ ,  $R^2=.67$ ,  $RMSD=1.45$ .

Contrast	Correlation	Coefficients	P-values
2	.38	.43	.03
3	-.40	-.18	.31
4	.16	-.19	.38
5	.15	.42	.02
6	.15	.04	.75
7	-.46	-.13	.38
8	.07	.02	.90
Prior Opinion	.25	.07	.64
Outcome Cost	-.63	-.40	.02

Table 7. Normalized means, correlations with burden of proof, standardized coefficients, and p-values for support strength ratings, prior opinion, and outcome cost that are regressed on burden of proof for Experiment 1. The overall regression was marginally significant,  $F(9,22)=6.87$ ,  $p<.001$ ,  $R^2=.74$ ,  $RMSD=1.29$ .

Claim	Mean Rating	Correlation	Coefficients	P-values
2	.46	.27	.08	.53
3	.94	-.49	-.39	.01
4	.77	.29	.20	.21
5	.86	-.11	.03	.82
6	.81	.37	-.02	.88
7	.93	-.42	-.22	.11
8	.70	.42	.32	.06
Prior Opinion	N/A	.25	.02	.87
Outcome Cost	N/A	-.63	-.37	.01

Table 8. Correlations with burden of proof, standardized coefficients, and p-values for support contrasts, prior opinion ratings, and outcome cost ratings that are regressed on burden of proof for Experiment 1. The overall regression was significant,  $F(9,22)=7.28$ ,  $p<.0001$ ,  $R^2=.75$ ,  $RMSD=1.26$ .

Contrast	Correlation	Coefficients	P-values
2	.29	.08	.51
3	-.47	-.45	.01
4	.26	.45	.01
5	-.10	.02	.86
6	.14	.08	.47
7	-.45	-.10	.45
8	.46	.34	.03
Prior Opinion	.25	.07	.57
Outcome Cost	-.63	-.40	.07

Table 9. Performance of models in fitting the burden of proof data.  $R^2$  and RMSD values for regressions using the specified model to fit the burden of proof data are reported. Add signifies Addendums, Conc signifies Concessions, All signifies a regression across experimental conditions, \*indicates  $p < .05$ ,  $^{\infty}$  indicates  $p < .01$ , and  $^s$  indicates  $p < .001$ .

	Experiment 1			Experiment 2		
	1Br	2Br	All	Add	Conc	All
Local Ratio	.11 1.97	.13 2.19	.12 2.03	.05 2.62	.22 2.21	.10 2.38
Support Ratio	.54 $^{\infty}$ 1.59	.64 $^s$ 1.26	.56 $^s$ 1.43	.14 2.50	.63 $^s$ 1.54	.31 $^s$ 2.09
Loc-Rel Ratio	.36 1.89	.34 1.70	.27 1.84	.23 2.34	.33 2.05	.28 2.14
Local Contrast	.13 2.20	.14 1.94	.13* 2.02	.07 2.60	.23* 2.21	.08 2.34
Local Contrast * Rel	.31* 1.74	.33* 1.92	.25 $^{\infty}$ 1.88	.24* 2.34	.31* 2.08	.28 $^{\infty}$ 2.14
Support Contrast	.56 $^{\infty}$ 1.74	.61 $^s$ 1.28	.58 $^s$ 1.58	.13 2.51	.58 $^s$ 1.63	.25 $^{\infty}$ 2.18
Support Contrast * Rel	.45 $^{\infty}$ 1.74	.63 $^s$ 1.28	.52 $^s$ 1.54	.19 2.43	.67 $^s$ 1.44	.38 $^s$ 1.98
Support SqContrast	.53 $^s$ 1.43	.60 $^s$ 1.49	.57 $^s$ 1.42	.11 2.53	.62 $^s$ 1.55	.31 $^s$ 2.10
Support SqCont * Rel	.48 $^{\infty}$ 1.71	.62 $^s$ 1.29	.54 $^s$ 1.47	.16 2.47	.67 $^s$ 1.46	.36 $^s$ 2.01

Table 10. Normalized Means, correlations with burden of proof, standardized coefficients, and p-values for local strength ratings, prior opinion ratings, and outcome cost ratings that are regressed on burden of proof for Experiment 2. The overall regression was significant,  $F(10,22)=3.87$ ,  $p<.01$ ,  $R^2=.65$ ,  $RMSE=1.78$ .

Claim	Mean Rating	Correlation	Coefficients	P-values
1	.47	-.25	.11	.51
2	.54	.42	.14	.45
3	.61	.25	.10	.52
4	.53	.32	.18	.32
5	.55	-.14	-.25	.10
6	.57	-.05	-.02	.87
7	.58	-.08	-.27	.15
8	.64	-.16	-.11	.51
Prior Opinion	N/A	.66	.66	.01
Outcome Cost	N/A	-.34	-.04	.81

Table 11. Correlations with burden of proof, standardized coefficients, and p-values for local contrasts, prior opinion ratings, and outcome cost ratings that are regressed on burden of proof for Experiment 2. The overall regression was significant,  $F(9,22)=4.48$ ,  $p<.001$ ,  $R^2=.65$ ,  $RMSD=1.74$ .

Contrast	Correlation	Coefficients	P-values
2	.26	.00	.99
3	-.51	-.42	.05
4	-.26	.05	.38
5	-.37	-.39	.04
6	.05	.21	.19
7	.19	.09	.50
8	.42	.08	.68
Prior Opinion	.66	.25	.26
Outcome Cost	-.34	-.20	.32

Table 12. Normalized means, correlations with burden of proof, standardized coefficients, and p-values for support strength ratings, prior opinion, and outcome cost that are regressed on burden of proof for Experiment 2. The overall regression was marginally significant,  $F(9,22)=4.62$ ,  $p<.001$ ,  $R^2=.65$ ,  $RMSD=1.73$ .

Claim	Mean Rating	Correlation	Coefficients	P-values
2	.66	-.14	-.29	.09
3	.83	-.56	-.13	.57
4	.57	.54	.26	.14
5	.84	-.55	-.16	.47
6	.78	.30	-.25	.13
7	.85	-.54	-.17	.51
8	.79	.32	.27	.25
Prior Opinion	N/A	.66	.23	.25
Outcome Cost	N/A	-.34	-.20	.20

Table 13 . Correlations with burden of proof, standardized coefficients, and p-values for support contrasts, prior opinion ratings, and outcome cost ratings that are regressed on burden of proof for Experiment 2. The overall regression was significant,  $F(9,22)=5.54$ ,  $p<.001$ ,  $R^2=.69$ ,  $RMSD=1.63$ .

Contrast	Correlation	Coefficients	P-values
2	-.03	-.37	.02
3	-.51	-.25	.18
4	.49	.19	.54
5	-.68	.34	.18
6	.39	-.12	.54
7	-.33	.08	.68
8	.58	.32	.20
Prior Opinion	.25	.14	.41
Outcome Cost	-.63	-.19	.27

Table 14. Percentage of burden of proof justifications by category for challenge structures and direct rebuttals in Experiment 3. Ap stands for anti-primacy, Str stands for claim strength, Sup stands for support, Agr stands for personal agreement, Evi stands for evidence, Rel stands for relevance, and Oth stands for other.

---

Structure	Justification Type						
	Ap	Str	Sup	Agr	Evi	Rel	Oth
Challenge Structure	3	3	24	26	27	7	10
Direct Rebuttal	4	4	23	24	28	6	11
Average	4	3	24	25	28	6	10

---

Table 15

Normalized means, standardized coefficients, and p-values for one-sided support strength ratings, prior opinion, and outcome cost that are regressed on burden of proof for Experiment 1. The overall regression was marginally significant,  $F(8,23)=5.67$ ,  $p<.001$ ,  $R^2=.66$ ,  $RMSE=1.43$ .

	Mean Rating	Coefficients	P-values
Claim			
3	.57	-.49	.01
4	.52	-.02	.87
5	.51	-.11	.43
6	.47	-.15	.41
7	.47	-.07	.59
8	.34	.22	.24
Prior Opinion	N/A	.02	.84
Outcome Cost	N/A	-.63	.01

Table 16

Normalized coefficients and p-values for one-sided support contrasts, prior opinion ratings, and outcome cost ratings that are regressed on burden of proof for Experiment 1.

The overall regression was significant,  $F(8,23)=3.80$ ,  $p<.01$ ,  $R^2=.57$ ,  $\text{RMSE}=1.62$ .

Contrast	Coefficients	P-values
3	-.37	.04
4	.06	.73
5	-.04	.77
6	.16	.31
7	-.14	.42
8	.02	.90
Prior Opinion	.09	.59
Outcome Cost	-.50	.01

## FIGURES

Figure 1. A graphical representation of argument 1. Arrows indicate clash relations between claims.

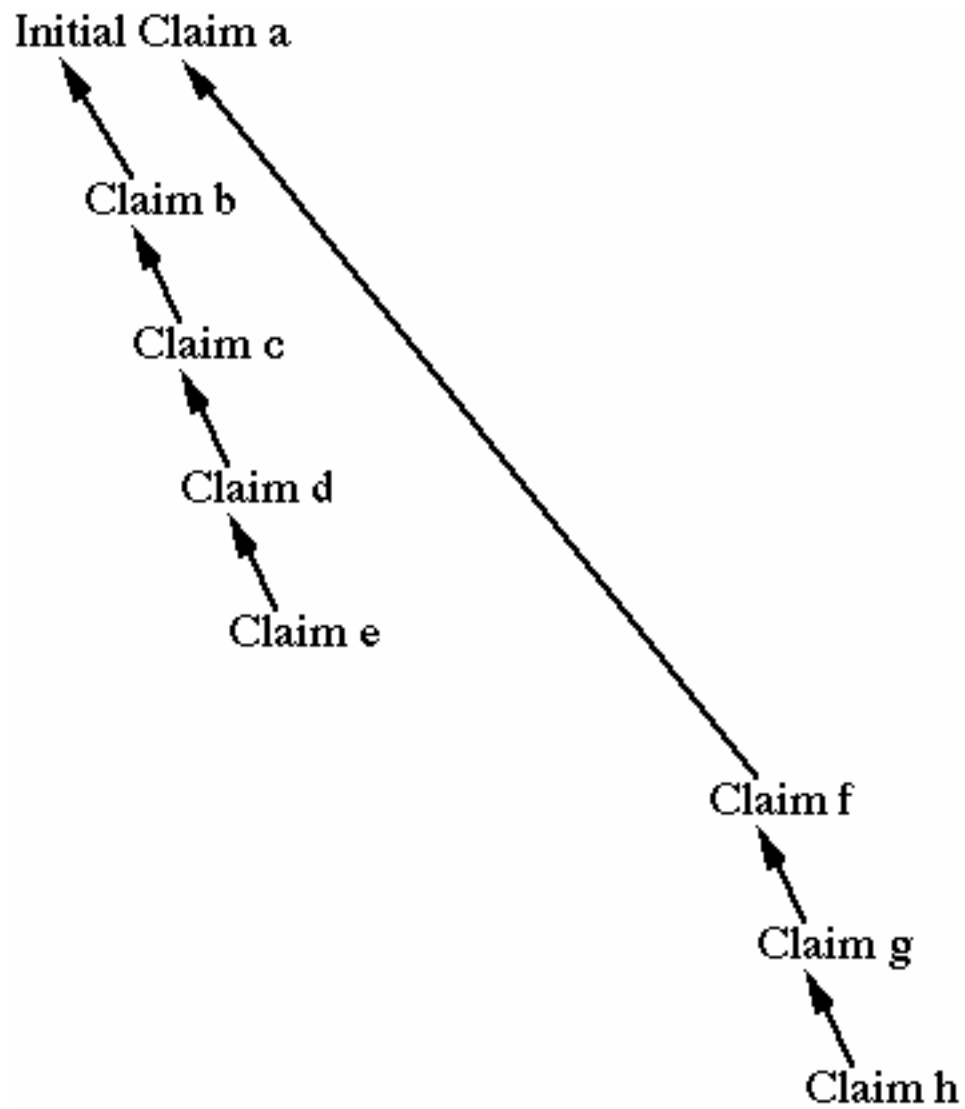
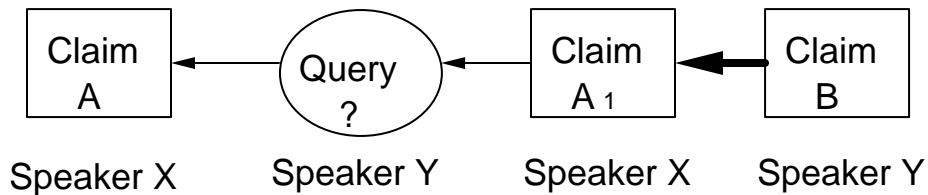


Figure 2. A graphical representation of the effect challenges have on argument structure. Bold lines represent contrast relations. Notice how speaker B has an extra contrast in the Challenge Version, even though the two speakers offer the same number of claims in both versions.

### Challenge Version



### Direct Rebuttal Version

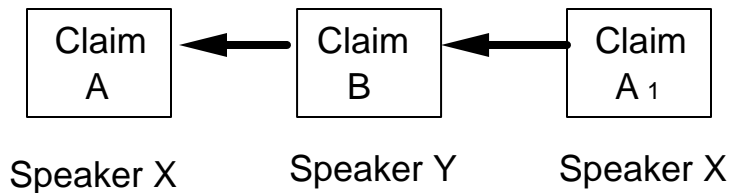


Figure 3. Results from Pilot Experiment 2. High values along the Y axis indicate high burden of proof for the first speaker while low values indicate high burden of proof for the second speaker.

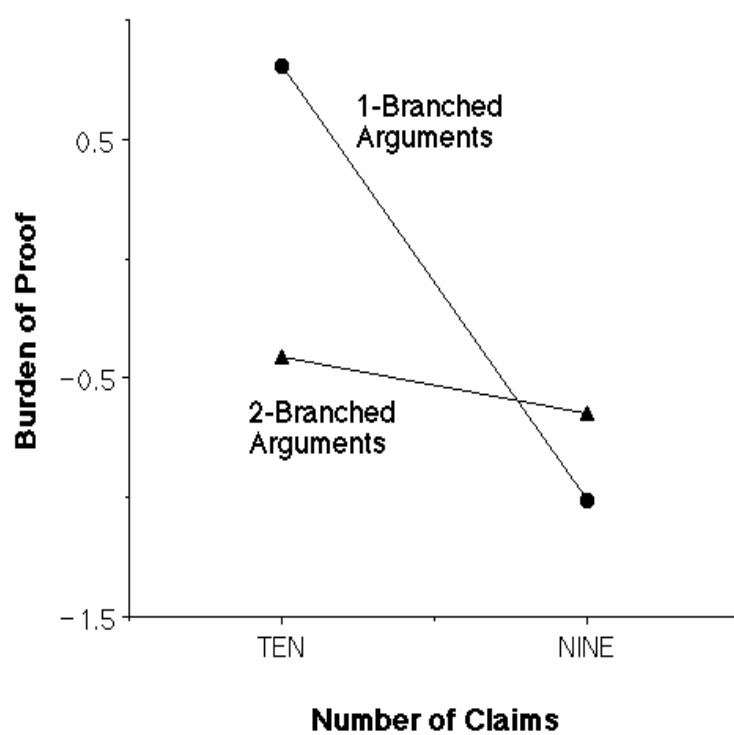


Figure 4. Average Local and Contextual Support strength Ratings from Experiment 1.

High values along the Y axis indicate higher strength.

Figure 5. Average Local and Contextual Support strength Ratings from Experiment 2.

High values along the Y axis indicate higher strength.

Figure 6. Average Local, Contextual Support, and One-Sided Support strength Ratings from Experiment 1. High values along the Y axis indicate higher strength.

## NOTES

1. This distinction is similar to that described by van Eemeren, Grootendorst, & Snoeck Henkemans (1996) between protagonist and antagonist.

2. The notion of staging can be traced back to Grimes (1975) in work done on determining how different components fit together in monologue text.

3. It is important to note that effect of coverage has recently been called into question by cross-cultural studies (Lopez, et al., 1997) as well as studies conducted on subjects who are experts in the topic of argument (Medin et al., 1997).

4. Note the similarity here between Kim and Pearl's representation and Toulmin's; target nodes roughly correspond to claims, data nodes correspond to data, and intervening nodes correspond to warrants.

5. This normalization assumes a valid interval scale. There are lots of reasons to suspect that this assumption may be violated (i.e., Parducci, 1982). However, for reasons of simplicity we overlook this potential bias in scaling.

6. The two arguments we eliminated had mean agreement ratings either over 5.55 or below 1.50.

7. For all of the analyses of this type, there are instances in which a claim (or contrast) has a coefficient with its sign in the wrong direction. However, on average, the coefficients for speaker 1's claims (and contrasts) are negative while the coefficients for speaker 2's claims (and contrasts) are positive.

## REFERENCES

- Ahn, W. & Bailenson, J. (1996). Mechanism-based explanations of causal attribution: An explanation of conjunction and discounting effect. Cognitive Psychology, 31, 82-123.
- Ahn, W., Kalish, C., Medin, D., & Gelman, S. (1995). The role of covariation versus mechanism information in causal attribution. Cognition, 54, 299-352.
- Allen, J.B. & Neely, S.T. (1997). Modeling the relation between the intensity just-noticeable difference and loudness for pure tones and wideband noise. Journal of the Acoustical Society of America, 102, 3628-3642.
- Bailenson, J. & Rips, L. J. (1996) Informal Reasoning and Burden of Proof, Applied Cognitive Psychology, 10, S13-S16.
- Bailenson, J. (1997). Claim Strength and Burden of Proof. Proceedings of the 19th Conference of the Cognitive Science Society, Palo Alto, Ca.
- Baron, J. (1991). Beliefs about thinking. In J. F. Voss, D.N. Perkins, & J. W. Segal (Eds.), Informal reasoning and education (pp. 169-186). Hillsdale, NJ: Erlbaum.
- Benassi, M. (1982). Effects of order presentation, primacy, and physical attractiveness on attributions of ability. Journal of Personality and Social Psychology, 46, 1230-1240.
- Birnbaum, M. H, Anderson, C. J., Hynan, L. G. (1989). Two operations for "ratios" and "differences" of distances on the mental map. Journal of Experimental Psychology: Human Perception & Performance, 15, 785-796.
- Brem, S. & Rips, L.J. (in press). Explanation and Evidence in Informal Argument, Cognitive Science.
- Carlson, R.A. & Dulaney, D. E. (1988). Diagnostic reasoning with circumstantial evidence. Cognitive Psychology, 20, 463.
- Cathcart, R.S. (1955). An experimental study of the relative effectiveness of four methods of presenting evidence. Speech Monographs, 22, 227-233.

Chaiken, S. (1979). Communicator physical attractiveness and persuasion. Journal of Personality and Social Psychology, 37, 1387-1397.

Clark, H. H. & Schaefer, E. F. (1989). Contributing to discourse. Cognitive Science, 13, 259-294.

Collins, A., & Michalski, R. (1989). A logic of plausible reasoning: a core theory. Cognitive Science, 13, 1-50.

Copper, C., Mullen, B., Asuncion, A., Gibbons, P., Goethals, G. R., Riordan, C., Schroeder, D., Tice, D., & Worth, L. (in press). Bias in the media: The subtle effects of a newscaster's smile. In B. Laczek (Ed.), Media effects.

Crano, W. D. (1977). Primacy versus recency in retention of information and opinion change. Journal of Social Psychology, 101, 87-96.

Cromwell, H. (1950). The relative effect on audience attitude of the first versus the second argumentative speech of a series. Speech Monographs, 17, 104-122.

Dunn, J. & Mann, P. (1987). Development of justification in disputes with mother and sibling, Developmental Psychology, 6, 791-798.

Drake, R. A. (1993). Processing persuasive arguments: 2. Discounting of truth and relevance as a function of agreement and manipulated activation asymmetry. Journal of Research in Personality, 27, 184-196.

Edwards, K. & Smith, E. E. (1996). A disconfirmation bias in the evaluation of arguments. Journal of Personality and Social Psychology, 71, 5-24.

Farley, A.M. and Freeman, K. 1995. "Burden of proof in legal argumentation", in Proceedings of Fifth International Conference on Artificial Intelligence and Law, 156-163.

Farley, A.M. and Freeman, K. 1995. "Dialectical Nonmonotonic Inheritance", in Proceedings of Sixth International Conference on Artificial Intelligence and Law.

Freeley, A. (1966). Argumentation and Debate. Belmont, Ca.: Wadsworth Publishing Co.

Furnham, A. (1986). The robustness of the recency effect: Studies using legal evidence. The Journal of General Psychology, 113, 351-357.

Garvey, C. (1987). Creation and avoidance of conflict in preschool children's play. Paper presented at the meeting of the Society for Research in Child Development, Baltimore.

Grice, H. P. (1989). Logic and conversation. In Studies in the way of words (pp. 1-143). Cambridge, MA: Harvard University Press.

Gulley, H. E. & Berlo, D. K. (1956). Effect of intercellular and intracellular speech structure on attitude change and learning. Speech Monographs, 29, 288-297.

Hamblin, C. L. (1970). Fallacies. London: Methuen.

Hogarth, R. M. & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. Cognitive Psychology, 24, 1-55.

Hovland, C.I., & Weiss, W. (1951). The influence of source-credibility on communication effectiveness. Public Opinion Quarterly, 15, 635-650.

Jackson, S., & Jacobs, S. (1980). The structure of conversational argument: Pragmatic bases for the enthymeme. Quarterly Journal of Speech, 66, 251-265.

Johnson J. T., Ogawa, K. H., Delforge, A. Early, D. (1989). Causal primacy and comparative fault: The effect of position in a causal chain on judgments of legal responsibility. Personality and Social Psychology Bulletin, 15, 161-174.

Johnson-Laird, P. N., & Byrne, R. M. J. (1991). Deduction. Hillsdale, NJ: Erlbaum.

Kahneman, D., Fredrickson, B.L., Schreiber, C. and Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end, Psychological Science, 4, 401-405.

Kassin, S. M. & Wrightsman, L.S. (1979). On the requirements of proof: The timing of judicial instruction and mock juror verdicts. Journal Personality and Social Psychology, 37, 1877-1887.

Kelman, H.C. & Hovland, C. I. (1953). Reinstatement" of the communicator in delayed measurement of opinion change. Journal of Abnormal and Social Psychology, 48, 327-335.

Kieras, D.E. (1978). Good and bad structure in simple paragraphs: Effects on apparent theme, reading time, and recall. Journal of Verbal Learning and Behavior, 17, 13-28.

Kim, J.H. & Pearl, J. (1987). CONVINCe: A conversational inference consolidation engine, IEEE Transactions on Systems, Man, and Cybernetics, 17, 120-132.

Knudson, R.E.(1994). An analysis of persuasive discourse: Learning how to take a stand, Discourse Processes. 18, 211-230.

Koehler, D. (1991). Explanation, imagination, and confidence in judgment. Psychological Bulletin, 110, 499-519.

Koehler, D. (1994). Hypothesis generation and confidence in judgment. Journal of Experimental Psychology: Learning, Memory & Cognition, 20, 461-469.

Koller, M. (1993). Rebutting accusations: When does it work, when does it fail? European Journal of Social Psychology, 23. 373.

Koszuth, A. (1991). Sexually abused child syndrome: Res ipsa loquitur and shifting the burden of proof. Law & Psychology Review, 15, 277-297.

Kuhn, D. (1997). Effects of dyadic interaction on argumentative reasoning. Cognition and Instruction. 15, 287-315.

Kuhn, D. (1991). The skills of argument. Cambridge: Cambridge University Press.

Kuhn, D., Flaton, R. & Weinstock, M. (1994) How well do jurors reason? competence dimensions of individual variation in a juror reasoning task. Psychological Science, 5, 289

Langer, E. J., & Roth, J. (1975). Heads I win, tails it's chance: The illusion of control as a function of the sequence of outcomes in a purely chance task. Journal of Personality and Social Psychology, 32, 951-955.

Lax, D.A. & Sebenius, J.K. (1991). Thinking Coalitionally: Party Arithmetic, Process Opportunism, and Strategic Sequencing. In, H. Peyton Young (Ed.), Negotiation Analysis (pp. 153-193). Ann Arbor, Mi: The University of Michigan Press.

Lewis, D. (1979). Score keeping in a language game. Journal of Philosophical Logic, 8, 339-359.

Linde, C. & Goguen, J. A. (1978). Structure of planning discourse. Journal of Social Biological Structure 1: 219-251.

Lopez, A., Atran, S., Coley, J.D., Medin, D.L., and Smith, E.E. (1997). The tree of life: Universal and cultural features of folkbiological taxonomies and inductions. Cognitive Psychology, 32, 251-295.

Lord, C. G., Ross, L., Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. Journal of Personality and Social Psychology, 37, 2098-2109.

Lowe, D. (1985). Co-operative structuring of information. International Journal of Man-Machine Studies, 22, pg. 97-111.

McConachy, R., Korb, K. B., & Zukerman, I. (1998). Deciding what not to say: An attentional-probabilistic approach to argument presentation. Proceedings of the Twentieth Annual Cognitive Science Conference, Madison, WI.

McGuire, W.J. (1985). Attitudes and attitude change. In G. Lindzey & E. Aronson (Eds.), Handbook of social psychology (vol. 2, pp. 233-346). New York: Random House.

McGuire, C. I. (1957). Order of presentation as a factor in "conditioning" persuasiveness. In C. I. Hovland (Ed.), The Order of Presentation in Persuasion (pp. 158-169). New Haven, CT: Yale Univ. Press.

Medin, D.L, Lynch, E.B., Coley, J. D. & Atran, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? Cognitive Psychology, 32, 49-96.

Morello, J.T. (1988). Argument and visual structuring in the 1984 Mondale-Reagan debates: The medium's influence on the perception of clash. Western Journal of Speech Communication. 52, 277.

Muntigl, P., & Turnbull, W. (1998). Conversational structure and face-work in arguing. Journal of Pragmatics, 29, 225-256.

Nickerson, R.S. (1991). Modes and models of informal reasoning: A commentary. In J. F. Voss, D.N. Perkins, & J.W. Segal (Eds.), Informal reasoning and education (pp.291-309). Hillsdale, NJ: Erlbaum.

Nisbett, R., & Ross, L. (1980). Human inference: Strategies and shortcomings of human judgment. Englewood Cliffs, NJ: Prentice-Hall.

O'Neill, B. (1991). Conflictual moves in bargaining: Warnings, threats, escalations, and ultimatums. In H. Peyton Young (Ed.), Negotiation Analysis (pp. 87-108). Ann Arbor, Mi: The University of Michigan Press.

Orsolini, M. & Pontecorvo, C. (1992). Children's talk in classroom discussions. Cognition and Instruction, 9, 113-136.

Orsolini, M. (1993). "Dwarfs do not shoot": An analysis of children's justifications, Cognition and Instruction, 11, pg. 281-297).

Osherson, D.N., Smith, E.E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. Psychological Review, 97, 185-200.

Parducci, A. (1982). Category ratings: Still more contextual effects. In B. Wegner (Ed.), Social attitudes and psychophysical measurement. Hillsdale, NJ: Erlbaum.

Pennington, N., & Hastie, R. (1993). Reasoning in explanation-based decision making. Cognition, 49, 123-163.

Pennington, N., & Hastie, R. (1986). Evidence evaluation in complex decision making. Journal of Personality and Social Psychology, 51, 242-258.

Petty, R. E., & Cacioppo, J.T. (1984). The effects of involvement on responses to argument quantity and quality: Central and peripheral routes to persuasion. Journal of Personality and Social Psychology, 46, 69-81.

Petty, R. E., & Cacioppo, J.T., & Goldman, R. (1981). Personal involvement as a determinant of argument-based persuasion. Journal of Personality and Social Psychology, 41, 847-855.

Petty, R. E., Schumann, D. W., Richman, S.A., & Strathman, A. J. (1993). Positive mood and persuasion: Different roles for affect under high- and low-elaboration conditions. Journal of Personality and Social Psychology, 64, 5-20.

Phillips, A. E. (1908). Effective Speaking. Chicago.

Pollock, J. (1989). How to build a person. Cambridge, MA: MIT Press.

Redelmeier, D.A., & Kahneman, D. (1996). Patient's memories of painful medical treatments: real-time and retrospective evaluations of two minimally invasive procedures. Pain, 66, 3-8.

Reichman-Adar, R. (1984). Extended person-machine interface, Artificial Intelligence, 22, 157-218.

Resnick, L. B., Salmon, M., Zeitz, C. M., Haley Wathen, S., Holowchak, M. (1993). Reasoning in Conversation, Cognition and Instruction, 11, 347-364.

Rips, L.J. (1998). Reasoning and Conversation, Psychological Review, 105, 411-441.

Rips, L.J. (1994). The psychology of proof: Deductive reasoning in human thinking. Cambridge, MA: MIT Press.

Rips, L.J. (1989). Similarity, typicality, and categorization. In S. Vosniadou and A. Ortony (eds.), Similarity and Analogical Reasoning. Cambridge University Press.

Romney, A.K., Weller, S.C., & Batchelder, W.H. (1986). Culture as consensus: A theory of culture and informant accuracy. American Anthropologist, 88, 318-338.

Scholten, A. (1991). Planning in ordinary conversation. Journal of Pragmatics, 16, 31-58.

Schonemann, P. H. & Lazarte, A. (1987). Psychophysical maps for subadditive dissimilarity ratings. Perception & Psychophysics, 42, 342-354

Shaw, V.F. & Johnson-Laird, P. N. (1993, November). Evaluating informal arguments. Poster session presented at the annual meeting of the Psychonomic Society.

Sillince, J. A. A. (1995). Shifts in focus and scope during argumentation. Journal of Pragmatics, 25, 413-431.

Sinclair, J. & Coulthard, M. (1975). Towards an analysis of discourse: The English used by teachers and students. Oxford: Oxford University Press.

Sperber, D. & Wilson, D. (1986). Relevance: Communication and Cognition. Cambridge: Harvard University Press.

Sponberg, H. (1946). A study of the relative effectiveness of climax and anti-climax order in and argumentative speech. Speech Monographs, 13, 33-44.

Slovan, S. A. (1994). When explanations compete: The role of explanatory coherence on judgments of likelihood. Cognition, 52, 1-21.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. Psychological Bulletin, *119* 3-22.

Stein, N. L., Bernas, S, Calicchia, D. J., Wright, A. (1996). Understanding and resolving arguments: The dynamics of negotiation. Models of understanding text. Lawrence Erlbaum Associates, Inc; Mahwah, NJ.

Tetlock, P. E. (1983). Accountability and the complexity of thought. Journal of Personality and Social Psychology, *27*, 501-526.

Thagard, P. (1992). Conceptual Revolutions, Princeton University Press; Princeton, NJ

Thagard, P. (1989). Explanatory Coherence, Behavioral and Brain Sciences, *12*, 435- 467.

Toulmin, S.E. (1958). The uses of argument. Cambridge, England: Cambridge Press.

Toulmin, S. E., Rieke, R., & Janik, A. (1979). An introduction to reasoning. New York: Macmillan.

Trognon, A. (1993). How does the process of interaction work when two interlocutors try to resolve a logical problem?, Cognition and Instruction, *11*, 325-345.

Van Wallendael , L. (1990). The quest for limits on noncomplementarity in opinion revision, Organizational Behavior & Human Decision Processes, *43* 385-405.

Vinokur, A. & Ajzen, I. (1982). Relative importance of prior and immediate events: A causal primacy effect. Journal of Personality and Social Psychology, *42*, 820-829.

Voss, J. F., Blais, J., Means, M. L., Greene, T. R., & Ahwesh, E. (1986). Informal reasoning and subject matter knowledge in the solving of economics problems by naive and novice individuals. Cognition and Instruction, *3*, 269-302.

Voss, J. W. & Means, M. L. (1991). Learning to reason via instruction in argumentation. Learning and Instruction, *1*, 337-350.

Walton, D. N.(1998). The New Dialectic: Conversational Contexts of Argument. Toronto: University of Toronto Press, Inc.

Walton, D. N. & Krabbe, E. C. (1995). Commitment in Dialogue. Albany, NY: SUNY Press.

Wangerin, P. (1993). A multidimensional analysis of the structure of persuasive arguments. Harvard Journal of Law and Public Policy, 16, 195-239.

White, K.G. & Wixted, J.T. (1999). Psychophysics of remembering. Journal of the Experimental Analysis of Behavior, 71, 91-113.

Wood, R. V. (1968). Strategic Debate. Skokie, IL: National Textbook Co.

Wigmore, J.H. (1935). A students' textbook of the law of evidence. Brooklyn: The Foundation Press.

Yates, J. F., & Curley, S. P. (1986). Contingency judgment: Primacy effects and attention decrement. Acta Psychologica, 62, 293-302.

## APPENDICES

Appendix I: Average Prior Agreement and Outcome Cost ratings used in Experiments 1 and 2. Subjects rated the topics on a 7 point scale; higher ratings indicate higher disagreement and a higher outcome cost.

First Claim in the Argument	Prior Agreement	Outcome Cost
“Recycling is a great benefit to society.”	1.97	5.00
“Evanston is a fun college town.”	5.13	2.03
“Tech needs a reading Week like CAS.”	4.00	3.59
“The El is the best form of transportation around here.”	3.57	2.00
“Carrying Mace won’t help in a dangerous situation.”	4.30	4.27
“The chances of a recent college graduate getting a job are pretty slim.”	5.09	4.40
“Communication is the most important factor in a relationship.”	1.63	4.69
“Giordano’s has the best pizza in Chicago.”	4.32	1.50
“The legal system needs to be reformed to prevent frivolous law suits.”	2.58	5.50
“We should raise tariffs on imported goods.”	4.47	5.72
“Capitol punishment executions should be televised in order to deter future crimes.”	5.53	5.69

“The government should spend less money on defense.”	3.38	6.00
“Students in Universities should be required to take physical education classes.”	4.85	2.80
“Burger World has got the best fast food burger in the city.”	5.02	1.41
“Modern remakes of old songs are usually terrible.”	4.00	1.84
“I think that more than just Democrats and Republicans should be allowed to participate in presidential debates.”	4.00	4.78

Appendix II: Relevance Instructions from Experiment 1 and Experiment 2.

“In the following experiment you will see a series of arguments. The arguments are conversations between two people who are discussing a certain issue and trying to persuade you to agree with what they are saying. Each person presents claims, or statements, in order to support their arguments. The speakers alternate turns such that each time a speaker makes a claim they are responding to the other speaker’s claim which came previously. Your job is to consider how relevant each claim is to the claim before it.

Sometimes a claim may be convincing in and of itself but it still is not very responsive to the claim that came before it. Consider the following claims:

1. Bill: We need to hire more policemen in Chicago for safety.
2. Jake: Five plus five equals ten.

In this example, Jake’s claim is very convincing in and of itself, since most people would agree that five plus five equals ten. However, clearly it is not very relevant to the issue of whether or not we need more policemen in the city. Therefore Jake’s claim is not very responsive to Bill’s.

This above example shows an extremely irrelevant claim to the one, which came before it. In arguments, usually claims are at least somewhat relevant. In the following example, Jake makes an argument which is not so drastically irrelevant as  $5 + 5 = 10$ , but still is not very responsive.

1. Bill: We need to hire more policemen in Chicago for safety.
2. Jake: Police officers usually work about 40 hours a week.

Here, while Jake's claim is relevant to the general topic of police officers, it still is not very responsive to Bill's claim. A claim can also be extremely responsive even if it is not convincing in and of itself.

Consider the following example.

1. Bill: We need to hire more policemen in Chicago for safety.
2. Jake: No we don't--there is no crime at all in Chicago.

In this example, Jake's claim clearly isn't very convincing, considering we all know that there is crime in Chicago. However, it still is relevant to the previous claim, since it directly addresses the topic of Bill's claim.

Your job is to read each claim in the argument and to decide how relevant it is to the claim which came before it. If it is relevant, then it does a good job in responding to the previous claim. Rate each claim on the following seven point scale:

not relevant      1.....2.....3.....4.....5.....6.....7      very relevant

If you think the claim is very responsive to the previous claim, then circle a seven. If it is not responsive at all, then circle a 1. For intermediate values, choose a number between the two. Remember, you are rating how responsive each claim is, not how convincing it is in and of itself. If you have any questions, please ask them now."

Appendix III: Support Strength Instructions from Experiment 1, Experiment 2, and Pilot Experiment 2.

“In the following experiment you will see a series of arguments. The arguments are conversations between two people who are discussing a certain issue and trying to persuade you to agree with what they are saying. In other words, each speaker has a position and is trying to convince you of that position. The speakers present claims, or statements, in order to support their arguments. Your job for each claim is to decide which speaker the claim provides support for, and how much support it provides. Consider the following example:

1. Jane: There are not enough policemen in Chicago.
2. Amy: There are plenty of cops--you see them all over the place.
3. Jane: Maybe there are a lot of cops around Evanston but in Chicago we need more.

In this example, each of the speakers has an argument. Jane argues that there are not enough policemen in Chicago, while Amy argues that there are enough policemen. Both Amy and Jane provide claims, which support their argument. Under each claim there is a ratings scale. Here is the same argument as the one above with the support rating scale.

1. Jane: There are not enough policemen in Chicago.
2. Amy: There are plenty of cops--you see them all over the place.

1.....	2.....	3.....	4.....	5.....	6.....	7
Strong	Moderate	Neutral	Moderate	Strong		
Jane Support	Jane Support			Amy Support	Amy Support	

3. Jane: Maybe there are a lot of cops around Evanston but in Chicago we need more.

1.....	2.....	3.....	4.....	5.....	6.....	7
Strong	Moderate	Neutral	Moderate	Moderate	Strong	
Jane Support	Jane Support			Amy Support	Amy Support	

4. Amy: But the Tribune today said that Chicago has more members on the police force than other cities its size.

1.....	2.....	3.....	4.....	5.....	6.....	7
Strong	Moderate	Neutral	Moderate	Moderate	Strong	
Jane Support	Jane Support			Amy Support	Amy Support	

If you think a claim provides strong support for Jane, then you should circle a 1 or a 2. The more support a claim provides for Jane's argument, the lower a number you should circle. If you think a claim provides strong support for Amy, then you should circle a 6 or a 7. The more support a claim provides for Amy's argument, the higher a number you should circle. If you think a claim does not provide support for either argument, then you should circle the number 4. Remember that you are circling numbers on the scale, not the words beneath them. If you have any questions, please ask them before you go on."

Appendix IV - Stimuli used in Experiment 3. The first eight arguments depict For-B arguments while the second eight depict Against-B arguments. The difference in the local strength of claims two and three is written above each argument, while the average local strength of each claim is written at the end of each line.

14-B(1.7)

1. Alan: Burger World has got the best fast food burger in the city. (3.13)
2. John: Burger World's burgers are horrible--the meat is tough and chewy. (4.74)
3. Alan: The patties used at Burger World are at least better than other fast food burgers. (3.04)

6-B(1.61)

1. Tom: The job market is very competitive (3.32)
2. Ron: Actually, lots of recent graduates are having success in the job market. (4.35)
3. Tom: The chances of a recent college graduate getting a job are pretty slim.(2.74)

4-B(2.46)

1. Bob: The El is the best form of transportation around here. (3.39)
2. Fran: Often times when I take the El I still have to get on a bus or walk for a while (6.85)
3. Bob: The El goes to more places in the city than any other transportation service. (4.39)

15-B(1.82)

1. Vince: Modern remakes of old songs are usually terrible. (2.91)
2. Ally: Some remakes are excellent--they take ideas from the past and bring them up to date in a current context. (5.39)
3. Vince: Some lyrics only make sense in the era in which they were composed. (3.57)

2-B(2.08)

1. Joan: Evanston is a fun college town. (3.5)
2. Mike: On Sherman Avenue there may be a few restaurants and bars, but certainly not enough to be considered a night life. (5.0)
3. Joan: There are lots of fun things to do at night on Sherman Avenue. (2.92)

8-B(1.05)

1. Bob: Giordano's has the best pizza in Chicago. (3.32)
2. Fran: Their pizza is not that great--they use low grade cheese that is oily. (5.05)
3. Bob: That's not true--they make incredibly tasty stuffed pizza. (4.0)

5-B(1.01)

1. Bill: Carrying mace won't help in a dangerous situation. (2.22)
2. Rick: Lot's of mace dispensers fix onto belts or key chains so they can visibly act as a deterrent. (5.71)
3. Bill: Not everyone who carries mace displays it so people can see it. (4.7)

11-B(1.26)

1. Elaine: Capitol punishment executions should be televised in order to deter future crimes. (3.27)
2. Billy: Just because someone is guilty of a certain crime does not mean they are not a human being. (6.0)
3. Elaine: Execution is reserved for situations where the most retributive punishment is in order--the criminal deserves no pity. (4.74)

7-A(2.82)

1. Sean: Communication is the most important factor in a relationship. (5.29)
2. Kary: Communication is not that important--there is no need to know every feeling of your partner (2.09)
3. Sean: But it is extremely important to talk and to relay feelings. (4.91)

3-A(1.20)

1. Julie: Tech needs a Reading Week like CAS. (4.09)

2. Dave: There is more lecture materials than readings--a Reading Week for Tech would shorten actual class term and the professors would not finish on time. (5.55)

3. Julie: My roommate is in Tech and she gets assigned tons of readings. (4.35)

16-A(1.76)

1. Calvin: I think that more than just Democrats and Republicans should be allowed to participate in presidential debates. (4.65)

2. Ronnie: If third party candidates voiced their opinions then people would realize that they are all a bunch of crackpots anyway. (2.91)

3. Calvin: Third parties have a chance to win--look at how well Ross Perot did in 1992. (4.65)

13-A(1.69)

1. Alice: Students in Universities should be required to take physical education classes. (2.83).

2. Roger: But students don't learn relevant skills from physical education classes. (2.82)

3. Alice: People develop their coordination, teamwork ability, and conditioning in gym classes. (4.51)

9-A(1.61)

1. Tim: The legal system needs to be reformed to prevent frivolous law suits. (4.83)

2. Bob: The legal system does not need to be reformed right now--there are ample resources. (2.82)

3. Tim: But there are thousands of unnecessary law suits raised each day. (4.43)

12-A(1.18)

1. Evan: The government should spend less money on defense.

2. Rich: Our current defense technologies are limited by lack of innovation, not lack of money--spending just will not help. (4.14)

3. Evan: But with current expenditures we can afford to hire innovative researchers to create new systems. (5.32)

1-A(1.93)

1. Michelle: Recycling is a great benefit to society. (4.78)

2. George: Actually, recycling is not all that beneficial. (1.73)

3. Michelle: It's a great method for reducing our waste. (3.65)

10-A(1.23)

1. Andy: We should raise tariffs on imported goods (2.83)

2. Erin: Well we can't just go and raise tariffs whenever we feel like it. (3.55)

3. Andy: Our own industries are struggling to get by because of cheap foreign products.(4.78)

## Appendix V – Additional Study on One-sided support ratings for Experiment 1.

In the study described in this section, I gathered support ratings for each side of the arguments used in Experiment 1 separately in order to distinguish between support for one's own claims and attacking opposing claims. Using these one-sided ratings to fit the contrast model, it becomes possible to demonstrate a more unique contribution of support.

### Method

Twenty-four subjects received packets consisting of an instruction page and all 16 argument topics. Subjects were volunteers recruited in Evanston and Chicago. Half the arguments were one-branch arguments and the other half were two-branch arguments. Of each branch type, half the arguments listed only the claims by speaker 1 and half the arguments listed only the claims by speaker 2. Consequently there were 4 arguments in each of the four (speaker by branch type) conditions. The packets were arranged such that both the orders of argument topics and conditions were randomized. Across subjects, each topic appeared in each condition an equal number of times.

Subjects rated how well each claim supported the speaker's argument as a whole. On each page there were four claims made by a single speaker. The claims were numbered one through four. For claims two, three, and four, subjects decided how much support each claim gave the speakers on a scale from one to seven, with higher numbers indicating more support for the speaker. Specifically, I asked them "how much support does each claim provide for the speaker's overall argument".

## Results

The mean ratings for each claim appear in Figure 6. Due to the different number of claims in each rating condition it was not possible to run a claim by rating condition ANOVA. However, an analysis with average rating as the dependent variable and rating condition as the independent variable demonstrated that the means were different,  $F(2, 61)=11.89$ ,  $p<.01$ . Furthermore, Tukey's HSD indicated that all three pairwise comparisons were significant with an alpha level of .05. One-sided support ratings were the lowest of the three ( $M = .49$ ), contextual support ratings were the highest ( $M = .78$ ) and local ratings were between the other two ( $M = .60$ ). As Figure X shows, one-sided ratings were the lowest in part due to extremely low ratings for claim 8. In fact, there was a significant linear trend by claim position,  $F(1,23)=8.74$ ,  $p<.01$ , with early claims rated higher than late claims in the argument. This effect may have occurred because as the argument progresses, late claims frequently interact with opposing claims. Because the one-sided support ratings were made without the presence of those opposing claims, the late claims didn't make as much sense as early claims to subjects.

I computed contrasts using the formulas from the model section, with the one-sided support ratings and relevance ratings as input. I also computed the same ratio baselines as in the main experiment (both with and without relevance ratings. I predicted that the contrast model using the one-sided support ratings would outperform the contrast model using local strength ratings. The one-sided ratings (which were made in the context of a single speaker's claims) provide more information about the argument

structure than the local ratings (which were not made in any type of argument context). Consequently they should more effectively predict burden of proof, which largely depends on argument structure.

However, the contrast model using the one-sided support ratings should not do as well as the contrast model using the contextual support ratings, since the latter model takes into account more information about opposing claims. The only information concerning opposing claims that the contrast model using one-sided support ratings utilizes comes from the relevance ratings (which only take into account a claim's relation to a single clashing claim). The contextual support ratings take into account a claim's relation with all opposing claims. Finally, I predicted that the contrast model using the one-sided support ratings would outperform the baselines.

The results partly demonstrate this pattern. When fitting the aggregate contrast model (with one-sided support ratings and relevance ratings) to the same 32 data points as described in Experiment 1, the model does not perform well,  $F(1,31)=10.43$ ,  $p<.01$ ,  $R^2=.26$ ,  $\text{RMSD}=1.86$ . In fact, the fit from the current analysis is quite similar to the fit of the contrast model using local strength ratings ( $R^2=.25$  in Table 9). Tables 15 and 16 show the results from regressions using the individual ratings and contrasts respectively. The overall  $R^2$ 's are a bit lower using the one-sided ratings than using the local strength ratings (Table 5) or local strength contrasts (Table 6), but these earlier analyses utilize more parameters in their analyses. The model slightly outperforms both the local-relevance ratio ( $F(1,31)=5.08$ ,  $p<.05$ ,  $R^2=.14$ ,  $\text{RMSD}=2.00$ ) and the ratio baseline without relevance ( $F(1,31)=7.51$ ,  $p<.01$ ,  $R^2=.20$ ,  $\text{RMSD}=1.94$ ).

In essence, the contrast model using contextual support ratings vastly outperformed the other two models (using local strength ratings and one-sided support ratings) because subjects making the contextual ratings had the advantage of viewing all opposing claims in the argument structure. As a result, those subjects could provide more information concerning burden of proof in their ratings than subjects from the other conditions, and the models using the contextual ratings account for twice as much variance in the burden of proof data as the other models. This is just a pilot study, however, and further research is necessary before drawing definitive conclusions.