When is Statistical Evidence Superior to Anecdotal Evidence in Supporting Probability Claims? The Role of Argument Type

Article *in* Human Communication Research · September 2009

DOI: 10.1111/j.1468-2958.2009.01360.x

CITATIONS
75

READS
3,339

2 authors:

Hans Hoeken
Utrecht University
210 PUBLICATIONS **2,394** CITATIONS

SEE PROFILE

Lettica Hustinx
Radboud University
35 PUBLICATIONS **331** CITATIONS

SEE PROFILE

HUMAN COMMUNICATION RESEARCH

Human Communication Research ISSN 0360-3989

ORIGINAL ARTICLE

When is Statistical Evidence Superior to Anecdotal Evidence in Supporting Probability Claims? The Role of Argument Type

Hans Hoeken[1] & Lettica Hustinx[2]

1 Department of Business Communication Studies, Centre for Language Studies, Radboud University Nijmegen, Nijmegen, The Netherlands
2 Department of Dutch Language and Culture, Centre for Language Studies, Radboud University Nijmegen, Nijmegen, The Netherlands

Under certain conditions, statistical evidence is more persuasive than anecdotal evidence in supporting a claim about the probability that a certain event will occur. In three experiments, it is shown that the type of argument is an important condition in this respect. If the evidence is part of an argument by generalization, statistical evidence is more persuasive compared with anecdotal evidence (Experiments 1 and 2). In the case of argument by analogy, statistical and anecdotal evidences are equally persuasive (Experiments 2 and 3). However, if the case in the anecdotal evidence is dissimilar from the case in the claim, statistical evidence is again more persuasive (Experiment 3). The implications of these results for the concept of argument quality are discussed.

doi:10.1111/j.1468-2958.2009.01360.x

Argument quality is believed to play an important role in the outcome of the persuasion process. According to dual-process models of the persuasion process, argument quality determines the outcome of the persuasion process if people are motivated and able to scrutinize the message, whereas it should have no effect when people lack the ability or motivation to do so (Chaiken, 1987; Petty & Cacioppo, 1986). However, argument quality also appears to have an impact on the attitudes of people who are less involved with an issue, and therefore less inclined to scrutinize the arguments. Park, Levine, Kingsley Westerman, Orfgen, and Foregger (2007), for instance, manipulated both the level of involvement and the quality of the arguments. Their results show that strong arguments had more impact on participants' attitudes than weak arguments regardless of the participants' level of involvement. The result

Corresponding author: Hans Hoeken; e-mail: h.hoeken@let.ru.nl

that argument quality influences the attitude of both highly and lowly involved participants has been obtained in other studies as well. In an (informal) meta-analysis, Park et al. showed that the main effect of argument quality was larger than the effect of the interaction between argument quality and involvement. Apparently, if the difference between strong and weak arguments is striking enough, even lowly involved participants are susceptible to this difference.

Given the theoretical and practical importance of argument quality, one would expect that the concept has been studied extensively. Fishbein and Ajzen (1981, p. 359) already put the issue on the agenda by stating that the lack of attention for how arguments influence the persuasion process "is probably the most serious problem in communication and persuasion research." Despite several calls to address the issue of what constitutes a strong argument (McGuire, 2000; O'Keefe, 2002; Reynolds & Reynolds, 2002), Johnson, Maio, and Smith-McLallen (2005, p. 654) conclude that there is (still) little research on the question of which factors make an argument strong. In many empirical studies, the development of strong and weak arguments is guided by the researchers' intuitions about what constitutes a strong argument, followed by a check to assess whether participants do share these intuitions. In this way, no independent criteria to distinguish strong from weak arguments are used. As a result of this approach, it remains unclear which aspects of the strong arguments make them strong. Therefore, the conclusion drawn by O'Keefe (2002, pp. 155–157) remains accurate: "identification of the active ingredients of argument quality is . . . a postponed matter."

A possible exception to the conclusion that we do not know which characteristics make an argument strong is the research in which the acceptance of a claim supported by statistics is compared with the acceptance of the same claim supported by an example. In this case, it seems reasonably safe to conclude that statistics are somewhat more convincing than examples (Reynolds & Reynolds, 2002). However, we will argue that this conclusion overlooks an important distinction developed within argumentation theory. According to argumentation theory, anecdotal evidence can serve as support for different types of argument. For one type of argument, statistics should indeed be more persuasive than examples, whereas for the other type, examples should be as convincing as statistics. After introducing the different argument types, three experiments are reported, which show the importance of taking argumentation theoretical distinctions into account when studying the relative persuasiveness of statistics compared with examples.

The relation between evidence types and argument types

Persuasive documents are often designed to influence people's behavior: To quit smoking, to drink less, or to buy a Volkswagen Beetle. The usually implicit claim is that the propagated behavior, for instance, to quit smoking, is better than its alternative (continue to smoke). Claims about the (relative) favorableness of an option are often supported by pragmatic argumentation (Van Eemeren, Grootendorst, &

Snoeck Henkemans, 1996, pp. 111–112). In this type of argumentation, the claim is supported by referring to the favorable consequences that acceptance of the claim will have. For instance, as a result of quitting smoking, it is argued that your chances of dying of lung cancer or of a heart attack are reduced, that your physical condition improves, and that your breath smells fresher. Crucial to the persuasiveness of this type of argumentation is that the audience regards the consequences as desirable as well as that it believes that the propagated behavior will indeed have these consequences. A strong argument in favor of a certain behavior would be that the behavior will probably result in a desirable consequence. A strong argument against a certain action would be that the behavior will probably result in an undesirable consequence.

The audience may doubt a consequence's desirability as well as its probability. In this case, evidence is needed to back up the claim. Areni and Lutz (1988) claim that assessing the desirability of a consequence is easier than assessing its probability. For instance, people may find it hard to assess the probability of lower heating costs as a result of installing a new heating system while they may find it easy to assess the desirability of such a cost reduction. Therefore, one would expect to encounter evidence in support of a probability claim more often than evidence in support of a desirability claim. Several studies, in which the use of evidence in public service documents is studied, indeed report that probability claims are far more frequently supported by evidence compared with desirability claims (Hornikx, Starren, & Hoeken, 2003; Schellens & Jong, 2004). The desirability of certain consequences (e.g., self-assurance, cheerfulness) is evident and no further information is needed to back this up. In this article, we focus on the evidence that can be provided to support probability claims.

Claims about the probability of a consequence can be supported by several types of evidence. Rieke and Sillars (1984, p. 91) distinguish between three types of evidence: Specific instances, statistics, and testimony. Rieke and Sillars (p. 92) regard as specific instances the ''giving of examples and illustrations.'' This type of evidence is also referred to as ''narrative evidence'' (Allen & Preiss, 1997; Feeley, Marshall, & Reinhart, 2006; Reynolds & Reynolds, 2002), ''story evidence'' (Baesler & Burgoon, 1994), and ''anecdotal evidence'' (Hoeken, 2001a; Koballa, 1986). Rieke and Sillars (p. 94) define statistics as ''a numerical compacting of specific instances.'' This type of evidence is referred to as statistical evidence by most authors (Allen & Preiss, 1997; Baesler & Burgoon, 1994; Feeley et al., 2006; Hoeken, 2001a; Reynolds & Reynolds, 2002). Testimony is described by Rieke and Sillars (p. 94) as citing a source to enhance the credibility of a certain claim and is usually referred to as testimonial evidence (Reinard, 1998).

In their review of the literature on evidence, Reynolds and Reynolds (2002) conclude that for evidence to enhance the acceptance of a claim, it has to be recognized as evidence, cognitively processed, and judged as legitimate. They note that many questions are yet unanswered, for instance, whether there are differences in the relative persuasiveness of different evidence types. They make an exception in

this respect for the comparison of statistical evidence versus anecdotal evidence. This question has been the focus of a relatively large number of experiments, enabling Allen and Preiss (1997) to conduct a meta-analysis on this issue. They conclude that ''the persuader is slightly more effective with a message that uses statistical proof as opposed to examples and narratives'' (Allen & Preiss, 1997, p. 128). Hornikx and Hoeken (2007) also report a persuasive advantage for statistical evidence compared with anecdotal evidence in a study in which participants rated the likelihood of a large number of consequences supported by either anecdotal or statistical evidence. O'Keefe (2002, p. 229) notes that there are large differences in the persuasiveness in some studies, but a lack of differences in others (Krupat, Smith, Leach, & Jackson, 1997). He concludes that ''a great deal is yet to be learned on this subject; the mechanisms underlying the various observed effects (and the conditions that encourage one or another effect) remain to be identified'' (O'Keefe, 2002, p. 229).

A promising condition for explaining when to expect a persuasive advantage of statistical evidence is the type of argument. From an argumentation theoretical point of view, the relative persuasiveness of statistical evidence compared with anecdotal evidence should depend on the type of argument the evidence is part of. Anecdotal evidence may serve as evidence in two different types of argument, namely, ''argument by generalization'' and ''argument by analogy.'' Rieke and Sillars (1984, p. 72) describe an argument by generalization as claiming *a general principle* from a series of instances. An example of such a general principle is the claim that ''juvenile delinquents grow up to become criminals.'' Baesler and Burgoon (1994) used this claim in their experiment on the relative persuasiveness of statistical and anecdotal evidences. The anecdotal evidence consisted of information about a single individual, the statistical evidence contained numerical information such as simple percentages and odds. From a normative point of view, the statistical evidence should be, and indeed was, more persuasive than the anecdotal evidence. Accepting a general claim about all juvenile delinquents growing up to become criminals on the basis of a single juvenile delinquent who did so is called the fallacy of hasty generalization. A more sound basis for accepting this claim would be to present a series of instances in which this was the case. The statistical information presented by Baesler and Burgoon, 80% of juvenile delinquents grow up to become criminals, did provide such information. This information proved to be more persuasive than the anecdotal evidence.

An argument by analogy is the result of comparing two situations, which are believed to have the same essential characteristics (Rieke & Sillars, 1984, p. 76). For such arguments, the claim contains a more specific situation or entity. An example of such a claim is that putting extra streetlights on the sidewalk in the town of X would result in a decrease of the number of burglaries. Hoeken (2001b) used this claim in his study on the relative persuasiveness of anecdotal and statistical evidence. The statistical evidence consisted of information that in the 48 towns that had put extra streetlights on the sidewalk, the number of burglaries had decreased by an average of 42%. The anecdotal evidence consisted of information that putting extra streetlights on the sidewalk in the town of Y had resulted in a decrease of the number

of burglaries by 42%. Towns X and Y were presented, and perceived, as very similar. Anecdotal and statistical evidences were equally successful in persuading participants to accept the claim compared with participants who received the claim without supporting evidence. Again, this result is the desired one from a normative point of view. Essential to the quality of the argument by analogy is the extent to which the case in the claim is similar to the case in the evidence. If the similarity between these cases is sufficiently high, anecdotal evidence should be as persuasive as statistical evidence.

Hypotheses

In this article, three experiments are reported in which the predictions from argumentation theory are tested. In the first experiment, the differential impact of anecdotal and statistical evidences within the context of the argument by generalization is studied. That is, the relative persuasiveness of the two evidence types in support of a claim about the probability of a certain effect for a large population is assessed. In the remainder of this article, we refer to this type of claim as a general claim. Following the norms of argumentation theory, the arguments containing statistical evidence should be stronger than the ones containing anecdotal evidence.

H1: Statistical evidence yields stronger acceptance of a general claim than anecdotal evidence.

In the second experiment, the relative persuasiveness of the two types of evidence is studied within the context of both the argument by generalization and the argument by analogy. Given that the number of instances is an important factor for the evaluation of the argument by generalization, but not for the argument by analogy, an interaction between type of argument and type of evidence is predicted.

H2: Statistical evidence yields a stronger acceptance of a general claim than anecdotal evidence, whereas no difference between the two types of evidence is found for the acceptance of a claim on a specific instance.

In the third experiment, it is studied whether the similarity between the case in the anecdotal evidence and the case in the claim influences the persuasiveness of the anecdotal evidence in the argument by analogy.

H3: Dissimilar anecdotal evidence in support of a claim on a specific instance yields less acceptance of this claim than either similar anecdotal evidence or statistical evidence.

The three experiments are designed and executed along similar lines. Participants received a large number of claims on different topics that were supported by different types of evidence or were not supported by evidence at all. A large number of claims were used to be able to generalize over messages, a desirable aspect of research on message effects (Jackson, O'Keefe, & Jacobs, 1988; O'Keefe, 2002, p. 175).

Furthermore, people traveling by train served as participants for the experiments, which led to heterogeneous participant samples.

Study 1

Method

Materials

The material consisted of 20 claims on different topics in which a relation was stated between the presence of an attribute and the occurrence of a consequence, for instance: ''Relaxation rooms in offices lead to a sharp decline in absenteeism due to illness in the Netherlands'' or ''Playing slow music instead of fast music will increase a supermarket's turnover.'' Anecdotal and statistical evidence as well as filler evidence were construed for each claim. The anecdotal evidence typically consisted of the experience of one individual, for instance: ''Thomas Kepers works in a large office in the Randstad conurbation. He has not had to call in sick since he started using the relaxation room on the second floor.'' The statistical evidence consisted of a numerical summary of large number of cases, for instance: ''From 1990 till 2002, a large-scale study was conducted on the effects of relaxation facilities at work. In companies offering such facilities, absenteeism due to illness occurred 24% less often.''

Apart from anecdotal and statistical evidences, other types of evidence were constructed to prevent participants from being confronted with anecdotal or statistical evidences only. The first type of filler evidence consisted of an explanation of why the action would have the predicted effect. The claim about the decrease in absenteeism as a result of the use of relaxation rooms, for instance, was supported by: ''Providing relaxation rooms at work implies that employees are able to relax from their demanding jobs for an hour a day. As a result, work-related pressure and stress decrease, which leads to better health.'' The second type of filler evidence consisted of the inclusion of an expert who simply restated the claim. For instance: ''Professor De Boer is a leading scholar in the field of leisure time and organizational behavior. In his book *Room for Relaxing* he claims that the provision of relaxation rooms decreases the number of people taking time off from work due to illness.'' The different types of evidence were kept approximately equally long.

Design

Each participant rated the probability of 20 different claims, four of which were supported by anecdotal evidence, four by statistical evidence, four by an explanation of why the consequence should occur, four by ascribing the claim to an expert, and four claims that were not supported by any evidence. The claims were presented in a random order. There were five different versions of the experimental booklet to ensure that all participants rated an equal number of claims for each condition, and that each claim in combination with each type of evidence was rated by an equal number of participants. To that end, a Latin square design was used: The first claim

was supported by anecdotal evidence in booklet 1, by statistical evidence in booklet 2, by an explanation in booklet 3, by an expert in booklet 4, and not supported by evidence in booklet 5. The second claim was supported by anecdotal evidence in booklet 5, by statistical evidence in booklet 1, by an explanation in booklet 2, by an expert in booklet 3, and not supported by evidence in booklet 4. The third claim was supported by anecdotal evidence in booklet 4, and so on. The order in which the claims were presented was the same in each of the booklets.

Instrumentation

The 20 claims (and evidence) were presented in an experimental booklet. After each short text, the claim was repeated and participants had to indicate how likely they regarded the consequence to occur on a 5-point scale ranging from "*very unlikely*" to "*very likely*." Using a single item to measure probability lowers the reliability of the measurement. However, in a pretest of the questionnaire in which probability was measured using three items, participants expressed their irritation at having to check three, almost identical, items over and over again. The reliability of this scale was good (Cronbach's $\alpha = .89$). To prevent participants from dropping out of the experiment because of irritation, it was decided to use a single item. Next, participants were asked to indicate their opinion on the comprehensibility of the reasoning on a 5-point scale ranging from "*very hard to understand*" to "*very easy to understand*." The latter question was only asked if the claim was supported by evidence. It was included to check whether certain types of evidence were considered more difficult than others. At the end of the booklet, participants were asked to answer questions about their age, sex, and highest level of education.

Participants

A total of 160 participants took part in the experiment. The number of women was slightly higher (84) than the number of men (76). Their ages ranged from 17 to 85, with a mean of 29.7. The level of education ranged from high school to a master's degree.

Procedure

The participants were railway passengers. They were asked whether they were willing to participate in a study in which they would be asked to give their opinion on several issues. Care was taken to approach passengers traveling to different destinations and at different times of the day. Approximately 65% agreed to take part in the experiment. The most frequent reasons for declining to participate were being busy with other things or almost having reached one's destination. If they agreed to participate, they randomly received an experimental booklet. After they had filled out the booklet, they were told about the study's goal and any remaining questions were answered. On average, filling out the booklet took about 15 minutes.

Statistical analysis

In this multiple message design, participants rated the probability of four claims supported by anecdotal evidence and four claims supported by statistical evidence. Furthermore, for each of the 20 claims, probability ratings were obtained when the claim was supported by anecdotal evidence and when supported by statistical evidence. The design therefore enables an assessment of whether effects would have been obtained with different participants as well as with different messages. The design is similar to the designs used in psycholinguistic experiments. The preferred way in psycholinguistics to assess whether significant differences could be generalized over participants as well as messages is the multistage decision rule by Forster and Dickinson (1976). They advise computing the F_1 and F_2 tests as well as the so-called *min F′*.

The F_1 test refers to the analysis of variance (ANOVA) in which the effect of evidence type (E) is tested by assuming that the participants (P) constitute a random factor and the evidence and claim combinations constitute a fixed factor ($F_1 = MS_E/MS_{P\times E}$). The test is conducted by running a within-participants ANOVA taking for each participant the average score of the four claims supported by anecdotal evidence and the average score of the four claims supported by statistical evidence. A significant effect implies that, had these same items been rated by a comparable sample of different participants, the difference would probably be obtained again. The F_2 test refers to the ANOVA in which the effect of evidence type is tested by assuming that the claims constitute a random factor, whereas the participants constitute a fixed factor ($F_2 = MS_E/MS_{CwE}$). The test is conducted by running a within-claims ANOVA taking for each claim the average score of the 32 participants who rated this claim when it was supported by anecdotal evidence and the average score of the 32 participants who rated this claim when it was supported by statistical evidence. A significant effect implies that the same difference would probably be obtained if 20 different claims plus evidence combinations had been rated by the same participants. The *min F′* is a test in which the mean square associated with the effect of evidence type is divided by the sum of the error terms used in the F_1 and in the F_2: $min\ F' = MS_E/(MS_{P\times E} + MS_{CwE})$. As such, it is a conservative test. If this effect is significant, the difference in probability ratings as a result of evidence types would probably be obtained again if other participants and other claim plus evidence combinations had been used. The H_0 should be rejected only if the *min F′* is significant or if both the F_1 and F_2 are significant. If one of these last two tests is not significant, the H_0 should be accepted (Forster & Dickinson, 1976, p. 141). This way of analyzing data is still a widely accepted method to analyze such designs in psycholinguistics (Dilley & McAuley, 2008; Post, Marslen-Wilson, Randall, & Tyler, 2008; Scheutz & Eberhard, 2008).

In this experiment, as in the other experiments reported in this article, we analyzed the raw scores as well as difference scores using this strategy. The difference scores were computed as follows: The probability of each claim without any supporting evidence was rated by 32 participants. For each claim, the average probability rating

Table 1 The Mean Increase in Probability Scores Compared with the Baseline and the Comprehensibility Ratings (1 = *Very Hard to Understand*, 5 = *Very Easy to Understand*) as a Function of Evidence Type for Study 1

	Baseline	Anecdotal Evidence	Statistical Evidence
Probability	2.87 (0.67)	$+0.13_a$ (0.72)	$+0.51_b$ (0.69)
Comprehensibility		3.91_a (0.90)	3.98_a (0.80)

Note: Different row subscripts indicate significant differences between conditions.

of these participants in the "without evidence" condition served as a baseline. The next step was to subtract this baseline score from each probability score for this same claim generated by a participant who read the claim either in combination with anecdotal evidence or with statistical evidence. The resulting difference scores were entered into the analysis. Because the same score was subtracted from the individual scores for the claim plus anecdotal evidence and the claim plus statistical evidence, the results of the ANOVAs taking the raw scores and the difference scores were identical. In the Results section, the difference scores are reported. They reveal the impact of providing anecdotal or statistical evidences on the probability ratings compared with the ratings of claims not supported by evidence. This strategy for data analysis and presentation is used in Experiments 2 and 3 as well.

Results

In the first column in Table 1, the average probability rating of claims without supporting evidence is displayed. The average rating is slightly below the neutral midpoint of the scale. If statistical evidence was used to support the claim, it was accepted more strongly than when anecdotal evidence was used [$F_1(1, 159) = 30.28,\ p < .001,\ \eta^2 = .16;\ F_2(1, 19) = 27.57,\ p < .001,\ \eta^2 = .59;$ *min* $F'(1, 63) = 14.43,\ p < .001$]. Both types of evidence were rated as equally comprehensible [$F_1(1, 159) = 1.43,\ p = .23;\ F_2(1, 19) = 1.70,\ p = .21$].

Discussion

As predicted by the first hypothesis, statistical evidence in support of a general claim was more persuasive than anecdotal evidence. This result could be interpreted as showing that participants apply the norms developed in argumentation theory: They are more reluctant to accept a claim over a class of cases that is supported by only one case, than to accept this claim when it is supported by a numerical summary of many cases. Another interpretation could be that statistics are considered more persuasive anyway. Reinard (1988, p. 23) remarks that "statistics have been respected in Western culture almost as icons of objectivity." In that case, statistical evidence would be more persuasive than anecdotal evidence regardless of whether it should be from a normative point of view.

To test this latter explanation, the second experiment was conducted. In this experiment, general claims on a class of cases were developed as well as specific claims

on a specific case. An example of a general claim would be that ''Elderly people would feel less isolated if they were to use the Internet,'' an example of a specific claim is: ''Sixty-nine-year-old Arnold de Beer would feel less isolated if he were to use the Internet.'' The combination of anecdotal evidence with a general claim yields an argument by generalization: The claim about the effect of using the Internet on the feelings of isolation for all elderly people is supported by reference to the effect it had for one person. For an argument by generalization, the number of cases in support of a claim is important, and therefore statistical evidence should be more convincing than anecdotal evidence. The combination of anecdotal evidence with a specific claim yields an argument by analogy. For this argument type, the similarity between the case in the claim and the case in the evidence is important, and therefore statistical evidence should not necessarily be more convincing than anecdotal evidence. If participants indeed follow the norms of argumentation theory, statistical evidence should be more persuasive than anecdotal evidence in support of a general claim, but no difference is expected when supporting a specific claim. If, however, statistical evidence derives its persuasive power from its aura of objectivity, then it will be more persuasive than anecdotal evidence regardless of the type of claim.

Study 2

Method

Materials

The material consisted of 24 claims on different topics in which a relation was stated between the presence of an attribute and the occurrence of a consequence. For each claim, a general and a more specific claim was constructed, for instance: ''Many elderly feel isolated from society. The use of the Internet can decrease this feeling of discomfort,'' and ''69-year-old Arnold de Beer feels isolated from society. Using the Internet can decrease this feeling of discomfort.'' Subsequently, for each claim, anecdotal and statistical evidences were constructed. The anecdotal evidence typically consisted of the experience of one individual, for instance: ''Since 72-year-old Bernhard van Delft has been online, he feels less lonely and cut off from the world around him.'' This anecdotal evidence was used to support the general as well as the specific claim. The statistical evidence consisted of a numerical summary of a large number of cases, for instance: ''Only 31% of the elderly with access to the Internet feels cut off from the world around them. For elderly without access to the Internet, this percentage is 64%.'' The different types of evidence were kept approximately equally long.

Design

Each participant rated the probability of 24 different claims, 12 general and 12 specific claims. Three general claims were supported by anecdotal evidence, three by statistical evidence, three by filler evidence consisting of an explanation why the consequence would occur, and three claims were not supported by any evidence; the same goes

for the specific claims. A Latin square design was used to ensure that each participant rated the probability of all 24 claims, whereas each claim in combination with each type of evidence was rated by an equal number of participants. This resulted in eight different versions of the experimental booklet. The order in which the claims were presented was identical in each of the versions.

Instrumentation

The 24 claims (and evidence) were presented in an experimental booklet. After each text, the (general or specific) claim was repeated and participants had to indicate how likely they regarded the consequence to occur on a 5-point scale ranging from "*very unlikely*" to "*very likely*." At the end of the booklet, participants were asked to answer questions about their age, sex, and highest level of education.

Participants

A total of 250 participants took part in the experiment. The number of women was slightly higher (131) than the number of men (119). Their ages ranged from 18 to 87, with a mean of 33.5 years. The level of education ranged from high school to a master's degree.

Procedure

As in Experiment 1, the participants were railway passengers. They were asked whether they were willing to participate in a study in which they would be asked to give their opinion on several issues. Approximately 70% of the passengers agreed to take part in the experiment, the reasons for declining being the same as in the previous experiment. If they agreed to participate, they randomly received an experimental booklet. After they had filled out the booklet, they were told about the study's goal and any remaining questions were answered. On average, filling out the booklet took about 15 minutes.

Results

Again, the probability ratings of the claims in both the argument by generalization and the argument by analogy were about neutral on a 5-point scale (Table 2). There was a main effect of evidence type revealing that statistical evidence was more persuasive than anecdotal evidence [$F_1(1, 249) = 10.85$, $p < .01$, $\eta^2 = .04$; $F_2(1, 23) = 7.58$, $p < .05$, $\eta^2 = .25$; *min* $F'(1, 87) = 5.26$, $p < .05$]. However, this main effect was qualified by the predicted interaction between argument type and evidence type. Although both the F_1 analysis [$F_1(1, 249) = 6.22$, $p = .01$, $\eta^2 = .02$] and the F_2 analysis [$F_2(1, 23) = 4.94$, $p < .05$, $\eta^2 = .18$] revealed significant effects, the (conservative) *min F*' did not reach conventional levels of significance [*min* $F'(1, 70) = 2.75$, $p = .10$]. Planned comparisons revealed that statistical evidence was more persuasive than anecdotal evidence when used in an argument by generalization, an effect that proved significant even if the *min* F' was used [*min* $F'(1, 54) = 9.91$, $p < .005$]. However, when used in an argument by analogy, statistical evidence was as persuasive as anecdotal evidence ($F_1 < 1, F_2 < 1$).

Table 2 The Mean Increase in Probability Scores Compared with the Baseline as a Function of Argument Type and Evidence Type for Study 2

	Baseline	Anecdotal Evidence	Statistical Evidence
Argument by generalization	3.03 (0.81)	$+0.17_{a}$ (0.68)	$+0.41_{b}$ (0.72)
Argument by analogy	2.97 (0.77)	$+0.19_{a}$ (0.73)	$+0.23_{a}$ (0.73)

Note: Different row subscripts indicate significant differences between conditions.

Discussion

The results of Experiment 2 show that the participants are sensitive to the type of argument that the anecdotal evidence is part of. In the context of an argument by generalization, the number of cases in the evidence is important, and, indeed, statistical evidence proved more persuasive than anecdotal evidence. In the context of an argument by analogy, the number of cases is less important than the similarity between the case in the evidence and the case in the claim. Although exactly the same anecdotal and statistical evidences were used in both types of argument, statistical evidence was only more persuasive if used in an argument by generalization.

The similarity between the case in the evidence and the one in the claim should be important to the persuasiveness of the anecdotal evidence in an argument by analogy. The similarity was not manipulated in Experiment 2 as care was taken to select cases that were similar to the one in the claim. To assess whether participants are sensitive to the similarity criterion, a third experiment was conducted. In this experiment, the extent to which the case in the evidence was similar to the case in the (specific) claim was manipulated. It was studied whether similar anecdotal evidence was indeed more persuasive than dissimilar anecdotal evidence in the context of an argument by analogy.

Study 3

Method

Materials

Sixteen claims on different specific cases were developed, for instance, the claim that extending the wine list would lead to more customers buying drinks in a diner in a midsized town in the southern part of the Netherlands: ''Diner 'het Hommeltje' in Heerlen is a successful diner where you can eat well at low prices. However, customers do not order many drinks. Extending the wine list at 'het Hommeltje' in Heerlen is a good option to increase the drinks turnover.'' Similar and dissimilar anecdotal evidences were developed. The similar evidence consisted of the successes of a similar type of restaurant in a similar town (''For in diner 'Den dikken dragonder' in Kerkrade, which aims at a similar clientele, extending the wine list has led to sharp increase in the drinks turnover.''), whereas the dissimilar evidence consisted of the successes of a Michelin-rated top restaurant (''For in top restaurant Da Vinci in Maasbracht, proud owner

of a Michelin star, extending the wine list has led to a sharp increase in the turnover of drinks.'') Apart from anecdotal evidence, statistical evidence was also developed for each claim, for instance: ''A study among 829 restaurants shows that restaurants with an extended wine list have a 23% higher turnover of drinks.''

Design

Each participant rated the probability of 16 different claims. Four claims were supported by similar anecdotal evidence, four by dissimilar anecdotal evidence, four by statistical evidence, and four claims were not supported by any evidence. Using a Latin square design ensured that each participant rated the probability of all 16 claims, whereas each claim in combination with each type of evidence was rated by an equal number of participants. This resulted in four different versions of the experimental booklet. The order in which the claims were presented was the same in each version.

Instrumentation

The 16 claims (and evidence) were presented in an experimental booklet. After each short text, the claim was repeated and participants had to indicate how likely they regarded the consequence to occur on a 7-point scale ranging from ''*very unlikely*'' to ''*very likely*.'' Because the comparison between the similar and dissimilar anecdotal evidence may give rise to subtle differences, it was decided to use a 7-point scale instead of a 5-point scale to be better able to assess such differences. After rating the 16 claims, the participants were asked to indicate the extent to which they regarded the case in the anecdote as similar to the case in the claim, to check whether the similarity manipulation had been successful. Each participant rated the similarity for the four similar anecdotes and for the four dissimilar anecdotes he or she had been confronted with. At the end of the booklet, the participants were asked to answer questions about their age, sex, and highest level of education.

Participants

A total of 88 participants took part in the experiment, 44 women and 44 men. Their ages ranged from 18 to 68, with a mean of 36.5 years. The level of education ranged from high school to a master's degree.

Procedure

As in the previous experiments, the participants were railway passengers who were asked to participate in a study about their opinions on several issues. Approximately 60% agreed to take part in the experiment. Those who agreed to participate randomly received an experimental booklet. After they had completed the booklet, they were told about the study's goal and any remaining questions were answered. On average, filling out the booklet took about 12 minutes.

Table 3 The Mean Increase in Probability Scores Compared with the Baseline as a Function of Evidence Type and the Mean Similarity Ratings (1 = *Very Dissimilar*, 7 = *Very Similar*) for Study 3

		Anecdotal Evidence		
	Baseline	Similar	Dissimilar	Statistical Evidence
Probability rating	3.99 (1.02)	$+0.68_a$ (0.99)	-0.04_b (1.16)	$+0.53_a$ (0.94)
Similarity rating		5.18_a (1.08)	2.53_b (1.15)	

Note: Different row subscripts indicate significant differences between conditions.

Results

In Table 3, the results of the experiment are displayed. The similarity manipulation proved successful: The similar anecdote was regarded as more similar to the case in the claim than the dissimilar anecdote [$F_1(1, 87) = 240.95$, $p < .001$, $\eta^2 = .74$; $F_2(1, 15) = 285.25$, $p < .001$, $\eta^2 = .95$; *min* $F'(1, 58) = 130.62$, $p < .001$]. The average probability rating of claims without supporting evidence was close to the neutral midpoint of the (7-point) scale. There was a significant effect of the different types of evidence [$F_1(2, 174) = 14.67$, $p < .001$, $\eta^2 = .14$; $F_2(2, 30) = 13.67$, $p < .001$, $\eta^2 = .48$; *Min* $F'(2, 95) = 7.23$, $p = .001$]. Planned comparisons revealed that, although the similar anecdotal evidence and the statistical evidence yielded higher probability ratings, supporting claims using dissimilar anecdotal evidence had no impact on claim acceptance at all.

Discussion

Although the participants regarded the similar anecdotal evidence as equally strong as the statistical evidence, they were much less convinced by the dissimilar anecdotal evidence. It appears that they grasp the idea that anecdotal evidence plays a different role in the context of an argument by analogy than in the context of an argument by generalization. Furthermore, they are aware of the fact that in the case of argument by analogy, the similarity between the case in the evidence and the case in the claim is of utmost importance. In the next section, the relevance of the three experiments for the research on persuasion and on argument quality is discussed.

General discussion

The distinction between the argument by generalization and the argument by analogy proved relevant to explain and predict the relative persuasiveness of statistical evidence compared with anecdotal evidence. Statistical evidence is more persuasive than anecdotal evidence within the context of an argument by generalization; anecdotal evidence proves to be as persuasive as statistical evidence within the context of an argument by analogy (as long as the case in the anecdotal evidence is similar to the case in the claim). Given the importance of argument quality for the theory of,

research on, and practice of persuasive communication, insights from argumentation theory should not be neglected when addressing the question of what factors make an argument strong.

The participants in this study acted as if they were fully versed in argumentation theory. They were more susceptible to the arguments that they normatively should be more susceptible to. This is even more remarkable as the arguments they were exposed to were used to support claims about the likelihood that a certain event will occur. Several studies have shown that people are better at detecting flaws in arguments supporting desirability claims than in arguments supporting probability claims. Areni and Lutz (1988) had students rate the strong and weak arguments developed by Petty and Cacioppo for experiments within the elaboration likelihood model (ELM) framework (see, for a review, Petty & Cacioppo, 1986) with respect to the probability and desirability of the consequences that were referred to in these arguments. They found that the weak arguments referred to less desirable consequences compared with the strong arguments, whereas the likelihood of the consequences' occurrence was the same for the strong and weak arguments. Van Enschot, Hustinx, and Hoeken (2003) had the same arguments evaluated by experts in argumentation theory, who found differences with respect to desirability as well as with respect to probability between the strong and weak arguments. Apparently, it takes training in argumentation theory to distinguish between strong and weak arguments in support of probability claims, whereas such expertise is not needed to distinguish between strong and weak arguments in support of desirability claims. Further evidence for the claim that it is easier to spot weaknesses in arguments in support of desirability claims is provided by Hustinx, Van Enschot, and Hoeken (2007). They manipulated independently the strength of arguments in support of the desirability and the probability of the consequences of implementing a comprehensive exam (the issue that is used in many studies within the ELM framework). The attitude toward the exam proved susceptible to the manipulation of the arguments in support of the consequences' desirability but not to the manipulation of the arguments in support of the consequences' probability.

A possible explanation for the fact that the participants in the experiments reported here were able to distinguish between strong and weak arguments may lie in the discourse context. In the experiments reported in this study, participants read texts consisting of only a few sentences. The presentation of arguments in isolation differs from the way in which the arguments were presented in the studies described above. In those cases, the arguments were part of a longer text containing information not directly relevant to the argument and the claim. This information may have consumed part of the attention people need to detect the flaws in a weak argument. In several studies, it has been shown that increasing the cognitive effort needed to process the argument, for example, by increasing syntactic complexity, led to a decreased ability to distinguish strong from weak arguments (Chebat, Gelinas-Chebat, Hombourger, & Woodside, 2003; Lowrey, 1998).

Hoeken and Hustinx (2007b) provide some evidence for the claim that additional information may inhibit people's ability to distinguish strong from weak arguments. They conducted an experiment in which they manipulated two factors within the context of an argument by analogy. First, they manipulated the level of similarity between the case in the anecdotal evidence and the case in the claim (as was done in Study 3 in this article). Second, they manipulated the amount of context by adding content to the (similar and dissimilar) anecdotal evidence in one condition, while leaving this content out in the other condition. Hoeken and Hustinx replicated the findings of Study 3: The similar anecdotal evidence led to a stronger acceptance of the claim compared with the dissimilar anecdotal evidence. This effect, however, was obtained only in the ''no additional content'' condition; when content was added, the persuasive advantage of the similar anecdotal evidence disappeared. This effect can be interpreted as evidence for the claim that the presence of additional content may cloud people's judgment of the quality of the arguments under consideration.

Slater and Rouner (1996) showed that the persuasive advantage of statistical evidence may depend on the extent to which the claim is congruent with the audience's own values. In their study, statistical evidence had a positive impact for participants whose values were in accordance with the message's claim (about the dangers of alcohol abuse), whereas anecdotal evidence had a positive impact for participants whose values were incongruent with the message's claim. De Wit, Das, and Vet (2008) report a similar effect of narrative evidence (as they call it) in heightening the risk perception and intention to obtain vaccination against the hepatitis B virus of men who were at risk for infection with this virus. They argue that the health message they used conveys potentially threatening information, and therefore the narrative evidence may have had a stronger effect in a similar way as it had in the Slater and Rouner study. The claims used in our study were of a general nature and did not contain personally relevant information. It would be interesting to see whether the statistical evidence would indeed be less persuasive if the claims were more personally threatening to the participants.

The anecdotal evidence used in this study consisted of a short description of a single case. Although this operationalization of anecdotal evidence is a valid one, there are other ways in which anecdotal evidence could be operationalized. First of all, the description of the case could be more elaborate. More elaborate descriptions of cases have been shown to influence perceptions of reality. Such cases are considered exemplars, and according to exemplification theory, these cases are considered representative of a more general state of affairs (Hoeken & Hustinx, 2007a; Zillmann, 1999, 2006; Zillmann & Brosius, 2000). Second, the number of exemplars is important, too. Brosius and Bathelt (1994) studied people's perception of, for example, the number of people who like apple wine. They provided participants with base rate information as well as with a number of exemplars consisting of interviews with individuals who expressed their like (or dislike) of apple wine. Brosius and Bathelt manipulated the extent to which the number of

interviewed people liking apple wine corresponded or diverged from the base rate information. For instance, if the base rate information stated that 80% of the people no longer liked apple wine, the number of interviewed individuals who expressed their dislike of apple wine was either four of five (corresponding) or one of five (diverging). Participants had to estimate the percentage of people disliking apple wine. The distribution of opinions expressed in the interviews had a stronger impact on this estimate compared with the base rate information. The effect that the distribution of exemplars has a stronger impact on frequency perceptions than base rate information has been documented in other studies and contexts as well (Fagerlin, Wang, & Ubel, 2005; Ubel, Jepson, & Baron, 2001; Zillmann, Perkins, & Sundar, 1992).

In none of the experiments in this study was anecdotal evidence found to be more persuasive than statistical evidence. It is important to note that this effect is restricted to the types of argument studied in these experiments. All arguments supported claims about the probability of a certain consequence. However, the quality of an argument does not rest solely on the probability of the consequence, but also on the consequence's desirability. As discussed in the introduction, the desirability of a consequence is often quite obvious. It does not take much convincing that cost reductions or better health are desirable. However, if the desirability of a consequence is not that obvious, evidence is needed to make people accept that the consequence is desirable (or undesirable). For this type of claim, anecdotal evidence may be more persuasive compared with statistical evidence because it may paint a clearer picture of the desirability aspects of the situation. The famous quote allegedly used by Stalin refers to this idea: "The death of one man is a tragedy, the death of millions is a statistic." It is an interesting question to assess whether anecdotal evidence is more persuasive compared with statistical evidence when supporting a not too obvious desirability claim.

The issues discussed above point out to some limitations of the results. The arguments were presented in short texts and the claims were about the occurrence of rather neutral events. The anecdotal evidence consisted of a short description of a single case. In the experiments, the order in which the different claims were presented was always the same. This may have led to confounding the variability between the different claims with an order effect. However, the order of the different types of evidence supporting the claims was manipulated. An additional limitation follows from using a single item to measure persuasiveness. Single-item measures may lead to more variance in the responses, making it harder to obtain reliable differences between conditions. However, within each condition, the effect of, for example, anecdotal evidence is measured repeatedly, yielding a more reliable measure of the persuasiveness of anecdotal evidence compared with measuring this effect only once on one item. Furthermore, in all experiments significant differences were obtained, which suggests that the lower reliability of the single-item measure did not prevent us from finding effects. The number of arguments that were rated by each participant enabled additional analyses to assess the generalizability of the differences for different

claims, which is an often called for but seldom realized feature of message design research (Jackson et al., 1988). Of course, the generalizability is restricted to claims (and evidence) of a similar, relatively neutral nature. Finally, the participant samples in the three studies are more heterogeneous than many of the studies relying on undergraduates as participants. The research therefore provides an interesting answer to O'Keefe's (2002) question of under which conditions statistical evidence makes an argument stronger. Consequently, it also contributes to the larger question of what factors make an argument strong.

Acknowledgments

This research was supported by a grant from NWO (360-80-020), the Dutch Research Foundation. We thank Hans Beentjes, Berna Hendriks, Margot van Mulken, Donna Rouner, and two anonymous reviewers for their insightful comments and advice on earlier versions of this article.

References

Allen, M., & Preiss, R. W. (1997). Comparing the persuasiveness of narrative and statistical evidence using meta-analysis. *Communication Research Reports*, **17**, 331–336.

Areni, C. S., & Lutz, R. J. (1988). The role of argument quality in the elaboration likelihood model. In M. J. Houston (Ed.), *Advances in consumer research* (pp. 197–203, Vol. 15). Provo, UT: Association for Communication Research.

Baesler, E. J., & Burgoon, J. K. (1994). The temporal effects of story and statistical evidence. *Communication Research*, **21**, 582–602.

Brosius, H.-B., & Bathelt, A. (1994). The utility of exemplars in persuasive communications. *Communication Research*, **21**, 48–78.

Chaiken, S. (1987). The heuristic model of persuasion. In M. P. Zanna, J. M. Olson, & C. P. Herman (Eds.), *Social influence: The Ontario symposium* (pp. 3–40, Vol. 5). Hillsdale, NJ: Erlbaum.

Chebat, J. C., Gelinas-Chebat, C., Hombourger, S., & Woodside, A. G. (2003). Testing consumers' motivation and linguistic ability as moderators of advertising readability. *Psychology & Marketing*, **20**, 599–624.

De Wit, J. B. F., Das, E., & Vet, R. (2008). What works best: Objective statistics or a personal testimonial? An assessment of the persuasive effects of different types of message evidence on risk perception. *Health Psychology*, **27**, 110–115.

Dilley, L. C., & McAuley, J. D. (2008). Distal and prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language*, **59**, 294–311.

Fagerlin, A., Wang, C., & Ubel, P. A. (2005). Reducing the influence of anecdotal reasoning on people's health case decisions: Is a picture worth a thousand statistics? *Medical Decision Making*, **25**, 398–405.

Feeley, T. H., Marshall, H. M., & Reinhart, A. M. (2006). Reactions to narrative and statistical written messages promoting organ donation. *Communication Reports*, **19**, 89–100.

Fishbein, M., & Ajzen, I. (1981). Acceptance, yielding, and impact: Cognitive processes in persuasion. In R. E. Petty, T. Ostrom, & T. Brock (Eds.), *Cognitive responses in persuasion* (pp. 339–359). Hillsdale, NJ: Erlbaum.

Forster, K. I., & Dickinson, R. G. (1976). More on the language-as-a-fixed-effect fallacy: Monte Carlo estimates of error rates for F1, F2, F' and Min F'. *Journal of Verbal Learning and Verbal Behavior*, **15**, 135–142.

Hoeken, H. (2001a). Anecdotal, statistical, and causal evidence: Their perceived and actual persuasiveness. *Argumentation*, **15**, 425–437.

Hoeken, H. (2001b). Convincing citizens: The role of argument quality. In D. Janssen & R. Neutelings (Eds.), *Reading and writing public documents* (pp. 147–169). Amsterdam/Philadelphia: Benjamins.

Hoeken, H., & Hustinx, L. (2007a). The impact of exemplars on responsibility stereotypes in fund-raising letters. *Communication Research*, **34**, 596–617.

Hoeken, H., & Hustinx, L. (2007b). The influence of additional information on the persuasiveness of flawed arguments by analogy. In F. H. van Eemeren, J. A. Blair, C. A. Willard, & B. Garssen (Eds.), *Proceedings of the sixth conference of the International Society for the Study of Argumentation* (pp. 625–630). Amsterdam: SicSat.

Hornikx, J., & Hoeken, H. (2007). Cultural differences in the persuasiveness of evidence types and evidence quality. *Communication Monographs*, **74**, 443–463.

Hornikx, J., Starren, M., & Hoeken, H. (2003). Cultural influence on the relative occurrence of evidence types. In F. H. van Eemeren, J. A. Blair, C. A. Willard, & A. F. Snoeck Henkemans (Eds.), *Proceedings of the fifth conference of the International Society for the Study of Argumentation* (pp. 531–536). Amsterdam: SicSat.

Hustinx, L., van Enschot, R., & Hoeken, H. (2007). Argument quality in the elaboration likelihood model: An empirical study of strong and weak arguments. In F. H. van Eemeren, J. A. Blair, C. A. Willard, & B. Garssen (Eds.), *Proceedings of the sixth conference of the International Society for the Study of Argumentation* (pp. 651–657). Amsterdam: SicSat.

Jackson, S., O'Keefe, D. J., & Jacobs, S. (1988). The search for reliable generalizations about messages: A comparison of research strategies. *Human Communication Research*, **15**, 127–142.

Johnson, B. T., Maio, G. R., & Smith-McLallen, A. (2005). Communication and attitude change: Causes, processes, and effects. In D. Albarracín, B. T. Johnson, & M. P. Zanna (Eds.), *The handbook of attitudes* (pp. 617–669). Mahwah, NJ: Erlbaum.

Koballa, T. R. (1986). Persuading teachers to reexamine the innovative elementary science programs of yesterday: The effect of anecdotal versus data-summary communications. *Journal of Research in Science Teaching*, **23**, 437–449.

Krupat, E., Smith, R. H., Leach, C. W., & Jackson, M. A. (1997). Generalizing from atypical cases. How general a tendency? *Basic and Applied Social Psychology*, **19**, 341–365.

Lowrey, T. M. (1998). The effects of syntactic complexity on advertising persuasiveness. *Journal of Consumer Psychology*, **7**, 187–206.

McGuire, W. J. (2000). Standing on the shoulders of ancients: Consumer research, persuasion, and figurative language. *Journal of Consumer Research*, **27**, 109–114.

O'Keefe, D. J. (2002). *Persuasion, theory & research* (2nd ed.). Thousand Oaks, CA: Sage.

Park, H. S., Levine, T. R., Kingsley Westerman, C. Y., Orfgen, T., & Foregger, S. (2007). The effects of argument quality and involvement type on attitude formation and attitude change: A test of dual-process and social-judgment predictions. *Human Communication Research*, **33**, 81–102.

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. Berlin: Springer.

Post, B., Marslen-Wilson, W. D., Randall, B., & Tyler, L. K. (2008). The processing of English regular inflections: Phonological cues to morphological structure. *Cognition*, **109**, 1–17.

Reinard, J. C. (1988). The empirical study of the persuasive effects of evidence. The status after 50 years of research. *Human Communication Research*, **15**, 3–59.

Reinard, J. C. (1998). The persuasive effects of testimony assertion evidence. In M. Allen & R. W. Preiss (Eds.), *Persuasion. Advances through meta-analysis* (pp. 69–96). Cresskill, NJ: Hampton.

Reynolds, R. A., & Reynolds, J. L. (2002). Evidence. In J. P. Dillard & M. Pfau (Eds.), *The persuasion handbook* (pp. 427–444). Thousand Oaks, CA: Sage.

Rieke, R. D., & Sillars, M. O. (1984). *Argumentation and the decision making process*. New York: Harper Collins.

Schellens, P. J., & Jong, M. D. T. (2004). Argumentation schemes in persuasive brochures. *Argumentation*, **18**, 295–323.

Scheutz, M., & Eberhard, K. M. (2008). Effects of morphosyntactic gender features in bilingual language processing. *Cognitive Science*, **28**, 559–588.

Slater, M., & Rouner, D. (1996). Value-affirmative and value-protective processing of alcohol education messages that include statistical evidence or anecdotes. *Communication Research*, **23**, 210–235.

Ubel, P. A., Jepson, C., & Baron, J. (2001). The inclusion of patient testimonials in decision aids: Effects on treatment choices. *Medical Decision Making*, **21**, 60–68.

Van Eemeren, F. H., Grootendorst, R., & Snoeck Henkemans, A. F. (1996). *Fundamentals of argumentation theory*. Mahwah, NJ: Erlbaum.

Van Enschot, R., Hustinx, L., & Hoeken, H. (2003). The concept of argument quality in the elaboration likelihood model. In F. H. van Eemeren, J. A. Blair, C. A. Willard, & A. F. Snoeck Henkemans (Eds.), *Anyone who has a view: Theoretical contributions to the study of argumentation* (pp. 319–335). Dordrecht: Kluwer.

Zillmann, D. (1999). Exemplification theory: Judging the whole by some of its parts. *Media Psychology*, **1**, 69–94.

Zillmann, D. (2006). Exemplification effects in the promotion of safety and health. *Journal of Communication*, **56**, S221–S237.

Zillmann, D., & Brosius, H.-B. (2000). *Exemplification in communication: The influence of case reports on the perception of issues*. Mahwah, NJ: Erlbaum.

Zillmann, D., Perkins, J. W., & Sundar, S. S. (1992). Impression-formation effects of printed news varying in descriptive precision and exemplifications. *Medienpsychologie: Zeitschrift für Individual- und Massenkommunikation*, **4**, 168–185.