AFRL-HE-WP-TR-2000-0102

UNITED STATES AIR FORCE RESEARCH LABORATORY

FOUNDATIONS FOR AN EMPIRICALLY DETERMINED SCALE OF TRUST IN AUTOMATED SYSTEMS

Jiun-Yin Jian
Ann M. Bisantz
Colin G. Drury
James Llinas

CENTER FOR MULTISOURCE INFORMATION FUSION
DEPARTMENT OF INDUSTRIAL ENGINEERING
STATE UNIVERSITY OF NEW YORK AT BUFFALO
BUFFALO NY 14260

FEBRUARY 1998

INTERIM REPORT FOR THE PERIOD APRIL 1996 TO FEBRUARY 1997

Human Effectiveness Directorate
Crew System Interface Division
2255 H Street
Wright-Patterson AFB OH 45433-7022

NOTICES

TECHNICAL REVIEW AND APPROVAL

AFRL-HE-WP-TR-2000-0102

This report has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public.

This technical report has been reviewed and is approved for publication.

FOR THE COMMANDER

MARIS M. VIKMANIS
Chief, Crew System Interface Division
Air Force Research Laboratory

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
February 1998

3. REPORT TYPE AND DATES COVERED
Interim Report: April 1996 to February 1997

4. TITLE AND SUBTITLE
Foundations for an Empirically Determined Scale of Trust in Automated Systems

5. FUNDING NUMBERS
C: F41624-94-D-6000
PE: 62202F
PR: 7184
TA: 10
WU: 46

6. AUTHOR(S)
Jiun-Yin Jian, Ann M. Bisantz, Colin G. Drury, James Llinas

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Center for Multisource Information Fusion
Department of Industrial Engineering
State University of New York at Buffalo
Buffalo, NY 14260

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research laboratory, Human Effectiveness Directorate
Crew System Interface Division
Air Force Materiel Command
Wright-Patterson AFB, OH 45433-7022

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-HE-WP-TR-2000-0102

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*
One component in the successful use of automated systems is the extent to which people trust the automation to perform effectively. In order to understand the relationship between trust in computerized systems and the use of those systems, we need to be able to effectively measure trust. Although questionnaires regarding trust have been used in prior studies, these questionnaires were theoretically rather than empirically generated and did not distinguish between three potentially different types of trust: human-human trust, human-machine trust, and trust in general. A three-phased experiment, comprising a word elicitation study, a questionnaire study, and a paired comparison study was performed, in order to better understand similarities and differences in the concepts of trust and distrust, and between the different types of trust. Results indicated that trust and distrust can be considered opposites, rather than comprising different concepts. Components of trust, in terms of words related to trust, were similar across the three types of trust. Results obtained from a cluster analysis were used to identify 12 potential factors of trust between people and automated systems. These 12 factors were then used to develop a proposed scale to measure trust in automation.

14. SUBJECT TERMS
trust scale, trust in automation

15. NUMBER OF PAGES
46

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UNL

This page left blank intentionally.

PREFACE

This effort was accomplished under Contract F41624-94-D-6000, Delivery Order 0007 for the Air Force Research Laboratory's Human Effectiveness Directorate, Crew System Interface Division, Information Analysis and Exploitation Branch (AFRL/HECA). It was completed for the prime contractor, Logicon Technical Services, Inc. (LTSI), Dayton Ohio, under Work Unit No. 71841046: "Crew Systems for Information Warfare." Mr. Don Monk was the Contract Monitor and Mr. Gilbert Kuperman was the Technical Monitor.

This page left blank intentionally.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

This page left blank intentionally.

INTRODUCTION

Automation has become increasingly common both in complex, technical systems (e.g., aircraft), and in everyday life (e.g., automobile cruise control). One component in the successful use of automated systems is how much people trust these systems to perform effectively. For instance, trust can affect how much people accept and rely upon increasingly automated systems (Sheridan, 1988). Trust plays a role in influencing operators' strategies toward the use of automation (Lee & Moray, 1994). For instance, pilots of advanced automation aircraft were less trusting of the automated aircraft than they were of less advanced aircraft, because they did not know whether or not the new technology was reliable and accurate (National Research Council, 1997).

In order to understand the relationship between trust in computerized systems and the use of those systems, we need to be able to measure trust effectively. Such a measurement tool would allow researchers or designers of computerized systems to better predict patterns of use of such systems, based on operators' assessment of trust. Previous research has investigated various methods for measuring trust. For example, research in social psychology has studied interpersonal relationships through the use of questionnaires. Larzelere and Huston (1980) used questionnaires to measure trust, in terms of benevolence and honesty, between partners. From these questionnaire surveys, several factors of trust were identified, including such concepts as predictability, reliability, and dependability. Additionally, researchers have concluded that the importance of these factors may be dynamic, changing over time as relationships develop. For instance, Rempel, Holmes, and Zanna (1985) established a hierarchical model of trust, and believed that certain factors of trust may change with time and increasing emotional investment.

Additionally, in human-machine systems research, scientists have investigated trust in computerized processes by using trust questionnaires. For example, Singh, Molloy, and Parasuraman (1993) developed a rating scale to measure people's potential for complacency, by investigating attitudes towards everyday automated devices such as automated teller machines. Lee and Moray (1994) and Muir and Moray (1996) examined operators' trust in automated systems in a simulated supervisory process control task and constructed subjective rating scales to evaluate participants' perceptions of the reliability and trustworthiness of the automated systems. Some of these questionnaires were based in part on those used in the social psychology research on trust. For example, Lerch and Prietula (1989) studied trust in problem solving advice and used self-reported measures to investigate two factors, predictability and dependability, which were previously identified by Rempel et al. (1985). Lerch and Prietula (1989) obtained confidence ratings of trust in the source of the advice by using questionnaires.

One assertion of these studies is that trust is a multi-dimensional concept. The definitions provided seem to capture different aspects of people's everyday usage of "trust" (Muir, 1987). Although the questionnaires are similar in that they have treated trust as a multi-dimensional concept, the factors of trust, and thus the attributes and descriptors included in the questionnaires, have been based on different theoretical notions of trust, depending on the theoretical orientation of the researcher. For example, Rempel et al. (1985) concluded that trust would progress in three stages over time from predictability, to dependability to faith. Muir and Moray (1996) extended these three factors, and developed an additive trust model that contained six components: predictability, dependability, faith, competence, responsibility, and reliability. Sheridan (1988) also suggested possible factors in trust, including reliability, robustness, familiarity, understandability, explication of intention, usefulness, and dependence.

Additionally, the questionnaires differ in that some are designed to measure trust in a particular person or system, while others measure a more general, non-directed propensity to be trusting. For example, Larzelere and Huston (1980) and Rempel et al. (1985) designed questionnaire items that measured trust in a specific individual (a romantic partner), and Lee and Moray (1996) asked questions specific to the control of an experimental system. In contrast, work by Singh et al. (1993) addressed a general potential for complacency by using questionnaire items about a variety of automated systems.

Given the current state of research on trust measurement, several assertions can be made. First, as noted above, the questionnaires used to measure trust have included items based on different theoretical notions of trust, and have not been based on an empirical analysis which attempted to uncover multiple components of trust. Second, previous studies have generally assumed that the concepts of trust and distrust were opposites. It could be that these concepts (trust and distrust) in fact encompass very different types of concepts or factors, as for example, do the concepts of comfort and discomfort (Zhang, Helander, & Drury, 1996).

Third, the previous studies have not explicitly evaluated how trust between human and automated systems differs from trust between humans, or for that matter, from trust in general. Although researchers in human-machine systems have employed concepts of trust from sociological studies, there is no empirical basis for necessarily assuming that concepts of human-machine trust are identical to trust between humans. Were such differentiated scales developed, they could provide a potentially more reliable and valid tool for assessing people's trust in automated, computerized systems.

Given this state of research, and the fact that it is important to be able to assess people's trust in systems that are becoming increasingly automated and computerized, we determined that it was necessary to conduct a study in order to provide an empirically based tool for assessing trust. Additionally, a goal was to identify potential similarities and differences among concepts of generalized trust, trust between people, and trust between human and automated systems.

METHOD

To address these issues, a three-phased experimental study was conducted of the concept of trust by an individual in another individual or system. The goal of these experiments was to explore the underlying factors comprising the concepts of trust, and to develop a potentially more reliable and valid tool for assessing people's trust in automated systems. The experiments are modeled after those conducted by Zhang et al. (1996) who developed a measurement scale for the similarly complex notion of comfort.

In the first phase, a word elicitation study, we collected various words related to concepts of trust and distrust. In the second phase, a questionnaire study, we investigated how closely each of these words was related to trust or distrust in order to evaluate whether or not trust and distrust were opposites or represented somewhat different concepts, and whether or not concepts of trust and distrust were similar for general trust, trust between people, and trust between humans and systems. The third phase was a paired comparison study, in which participants rated the similarity of pairs of words. Data from both the questionnaire study and the paired comparison study were then used to construct a multi-dimensional measurement scale for trust.

EXPERIMENT 1: WORD ELICITATION STUDY

The objective of this phase was to collect a large set of words related to trust and distrust.

Method

Participants

Seven graduate students majoring in Linguistics or English were recruited, because of their presumed knowledge of word meanings. All participants were native English speakers; two were male and five were female. Participants were paid five dollars to complete one questionnaire. It took participants from 20 to 30 minutes to complete the task.

Procedure

There were three conditions in this experiment. Participants were asked to provide written descriptions of their understanding of both trust and distrust with respect to either trust between people, trust in automation, or trust with no specific qualification. Next, participants were also asked to rate whether a set of 138 words were related to trust using a nominal scale, with "positively related to trust," "not related to trust," "negatively related to trust," and "don't know" as scale points. This initial set of 138 words was collected by analyzing questionnaires used in previous studies, and from dictionary definitions and thesauri. As with the written descriptions, these ratings were performed with respect to the three conditions of trust between people, trust in automation, and general trust.

Results

We obtained 38 new words from the written descriptions of trust provided by the participants' questionnaires. In addition, we eliminated words from the initial set based on the participants' ratings of the words. Words which were rated "not-related to trust" by four or more or the seven participants and in all three contexts were eliminated. We also eliminated words that were ambiguous: that is, words which some participants rated as "positively related to trust" while other participants rated as "negatively related to trust." For example, the word "assertion" was judged to be both positively and negatively related to trust. These words may be ambiguously related to trust because their meanings are context dependent. To provide continuity with the existing literature, words retrieved from questionnaires used in previous research were not eliminated, although some were rated as "not related to trust" (e.g., familiarity). A total of 60 words were eliminated. The 60 eliminated words are shown in Table 1, and have an "x" in the Eliminated column. After eliminating these words and adding the new words, the final set of words, that we will refer to as Set-1, contained 96 trust-related words. Words in this set are shown in bold-faced type in Table 1 and were used in the subsequent questionnaire study.

Table 1. Word List. Words Shown in Bold Were Used in the Subsequent Questionnaire Study

Word	*Initial set*	*Added*	*Eliminated*
Absolute	X		X
Ambiguity	X		
Anger		X	
Aplomb	X		X
Apprehensive	X		
Assertion	X		X
Assurance	X		
Attack	X		
Bashfulness	X		X
Belief	X		
Believe	X		X
Benevolence	X		
Betray		X	
Beware		X	
Biased	X		
Bind	X		X
Can be relied upon	X		X
Cartel	X		X
Casual	X		X
Cautious	X		X
Caution		X	
Certain	X		X
Certainty	X		
Certitude	X		X
Charge	X		
Cheat	X		
Closeness		X	
Coalition	X		X
Commit	X		
Competence		X	
Complacency	X		
Confidence	X		
Confidential	X		X
Constancy	X		
Contingent	X		X
Conviction	X		
Cooperation		X	
Count on	X		X
Couple	X		
Courage	X		
Credence	X		X
Credit	X		
Creed	X		X
Cruel		X	
Custody	X		
Deception		X	
Declaration	X		
Definite	X		
Denial	X		
Denomination	X		X
Dependence	X		
Determination	X		
Diffidence	X		X
Disbelief	X		X
Dispute	X		
Distance		X	
Distrust		X	
Doctrine	X		X
Doubt	X		
Doubtless	X		X
Embody	X		X
End		X	
Entrust	X		
Error		X	
Faction	X		X
Faith	X		
Failure		X	
Familiarity	X		
Falsity		X	
Feeling	X		
Fidelity	X		
Firm	X		X
Firmness	X		X
Fixed	X		X
Friendship		X	
Fund		X	
Guardianship	X		
Harm		X	
Heresy	X		X
Hesitation	X		
Honesty	X		
Honor		X	
Inarguable	X		
Incontestable	X		
Incontrovertible	X		
Incorporate	X		X
Incredulity	X		X
Independence	X		X
Indisputable	X		X
Indubitable	X		X
Inducement	X		
Inevitable	X		X
Infidelity	X		X
Integrate	X		X
Integrity	X		
Intimacy		X	
Irrefutable	X		
Join	X		X
League	X		X
Lie		X	
Love	X		
Loss		X	
Loyalty	X		
Merge	X		X
Mingle	X		X
Misleading		X	
Mistake		X	

(Table 1 cont'd.)

Word	*Initial set*	*Added*	*Eliminated*
Mistrust	X		
Mix	X		X
Modesty	X		X
Moral		X	
Must have motivational relevance	X		X
Mutuality		X	
Naive		X	
Nobility		X	
Obligation	X		
Opinion	X		
Overcharge	X		
Overtrust		X	
Persistence	X		
Persuasion	X		
Phony		X	
Pledge	X		
Positive	X		
Predictability	X		
Principle	X		
Probable	X		X
Promise	X		
Question	X		
Reciprocate with fairness	X		X
Regret		X	
Reliability	X		
Reliable	X		X
Reliance	X		X
Rely on	X		X
Respect		X	
Respectful	X		X
Responsibility		X	
Robustness	X		
Scruple	X		
Secure	X		X
Security		X	
Security in caring response	X		X
Selfish	X		
Shyness	X		X
Sincere	X		
Timid	X		X
Trustworthy	X		
Unbelief	X		X
Undeniable	X		X
Understandability	X		
Unerring	X		
Unfailing	X		X
Unquestionable	X		X
Usefulness	X		
View	X		X
Wariness		X	
Wrong		X	
Yoke	X		X

EXPERIMENT 2: QUESTIONNAIRE STUDY

The objectives of the questionnaire study were to identify a smaller set of words related to trust and distrust for use in the next phase of the experiment, the paired-comparison phase. Paired-comparison studies are lengthy and tedious, and thus demand a relatively small word set. Additionally, the questionnaire study allowed us to evaluate two questions: first, to determine whether the concepts of trust and distrust are negatively related; and second, to determine whether concepts of trust and distrust are similar across general trust, trust between people, and trust between people and automated systems.

Method

Participants

One hundred-twenty participants were recruited from members of the university community. There were 45 graduate students and 75 undergraduate students, of whom 50 were male and 70 were female. All participants were native English speakers. Participants were paid five dollars to complete one questionnaire. It took participants from 20 to 30 minutes to complete the task.

Procedure

In this experiment, participants were asked to rate the extent to which words from Set-1 were related to trust or distrust, from the perspective of either trust in general, or trust between people, or trust in automated systems, for a total of six between-subject conditions. Participants rated the relatedness of the word to trust or distrust using a seven point scale, with end points of "positively related to trust (or distrust)" and "negatively related to trust (or distrust)."

Results

Participants' ratings were analyzed in several ways. First, for each word, average ratings of trust were correlated with average ratings of distrust, for each of the three conditions (general trust, human-human trust, and human-machine trust). Ratings of trust were highly negatively correlated with ratings of distrust ($r = -.96$, $r = -.95$, $r = -.95$, respectively). Thus, words that had a high positive rating for trust also had a high negative rating for distrust. This indicates that concepts of trust and distrust are in fact opposites, rather than comprising different factors. If any other factors are present, they can explain a maximum of 10% ($1-0.95^2$) of the variance in trust ratings.

A regression analysis was also performed: ratings of distrust were analyzed as a function of ratings of trust. Figure 1 shows the regression analysis across the three conditions. After comparing the slopes across general trust vs. human-human trust, general trust vs. human-machine trust, and human-human trust vs. human-machines trust, we found that there was no significant differences between general trust (slope = -0.96) and human-machine trust (slope = -1.01) ($t = 1.16$, $df = 220$). However, there were significant differences between general trust and human-human trust (slope = -0.79) ($t = 4.78$, $df = 220$), and human-human trust and human-machine trust ($t = 5.68$, $df = 220$). The slope of the line indicates that people were less extreme in their ratings of human-human distrust than trust. That is, a word would have a greater trust rating than a negative distrust rating, or a greater negative trust rating than distrust rating. This was not true for ratings of human-machine or general trust. These results seem to indicate that people might perceive trust and distrust with respect to human-human relationships slightly differently. This could be due to participants being more comfortable considering these relationships in terms of trust, rather than distrust, perhaps because an assessment of distrust in people seems more negative and unpleasant than an assessment of low or negative trust.

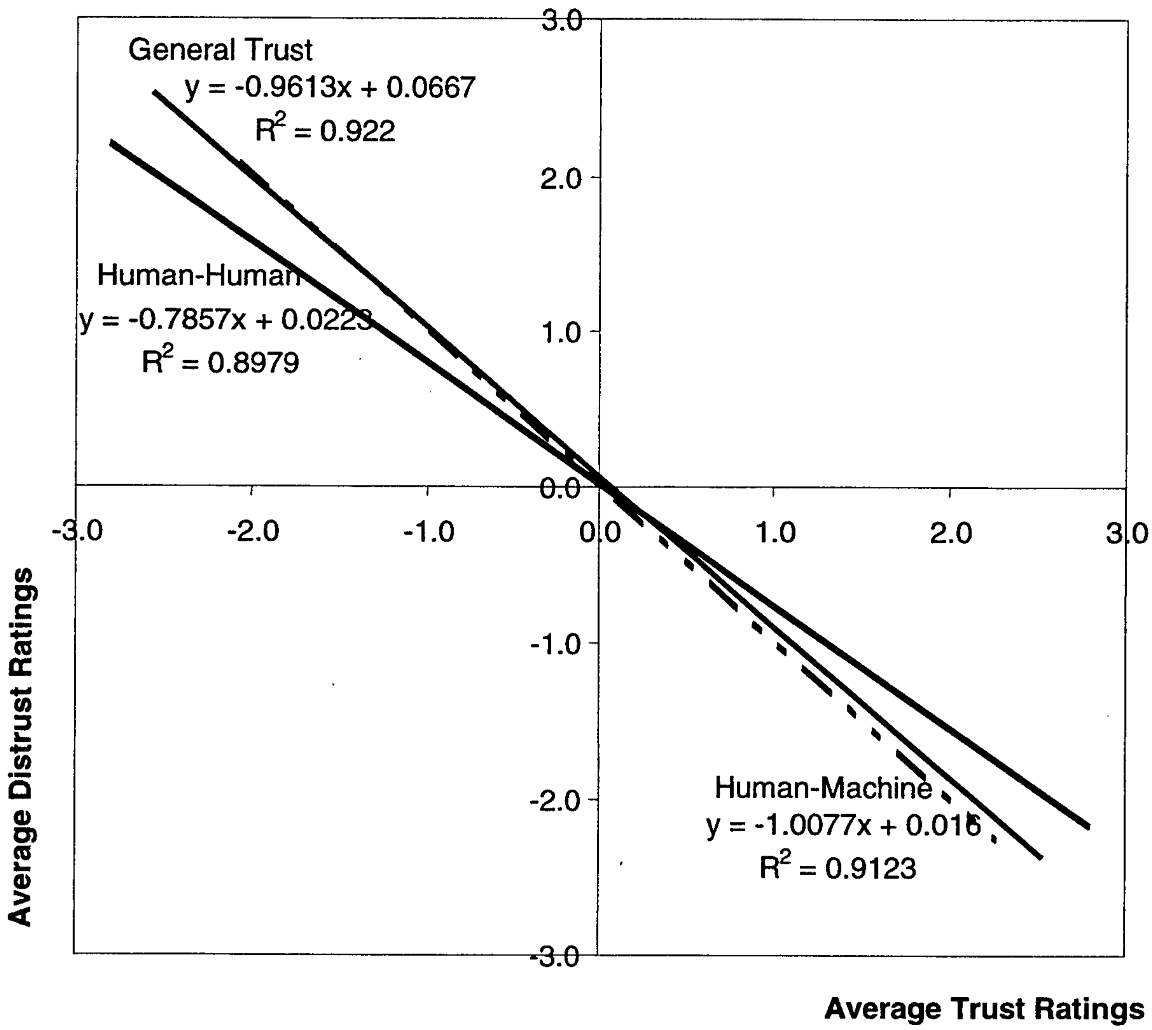

Figure 1. Regression analysis across three conditions.

Additionally, we compared ratings of individual words across the three conditions of general, human-human, and human-machine trust, to see how individual words might be differently related to the three types of trust. Words were assigned, according to their average ratings, into the top 5, 10, 15, 20, 25, and 30 words most related to trust and distrust, for each condition. For example, the five words most related to general trust were *trustworthy*, *honesty*, *loyalty*, *reliability*, and *honor*. The five words most related to trust between humans and automated systems were *trustworthy*, *loyalty*, *reliability*, *honor*, and *familiarity*. The five words most related to trust between people were *trustworthy*, *honesty*, *loyalty*, *reliability*, and *integrity*. The degree to which these sets overlap gives an indication of the extent to which concepts of trust and distrust were similar for the three conditions.

One measure of this overlap is the size of the union of the sets across the three conditions. For example, if the "top 5" sets for each condition were identical, than the union set size would be 5, indicating the highest possible similarity. If the "top 5" sets were completely different, the union set size would be 15, indicating no similarity across groups. For the "top 5" set then, the minimum union set size would be 5, while the maximum union set size would be 15. Continuing our example, Table 2 shows the top five words related to trust for each condition. The words *trustworthy, loyalty*, and *reliability* were common to all, giving an intersection of size three. Across the three conditions, the top five groups comprised a total of seven different words, giving a union of size seven.

Table 2. Five Words Most Related to Trust Across Three Conditions: Three Words in Common Give a Union Set Size of Seven

Conditions	General trust	Trust between people	Trust between human and automated systems
Words	1. Trustworthy	Trustworthy	Trustworthy
	2. Honesty	Honesty	
	3. Loyalty	Loyalty	Loyalty
	4. Reliability	Reliability	Reliability
	5. Honor		Honor
		6. Integrity	
			7. Familiarity

Union sets were determined for the top 5, 10, 15, 20, 25, and 30 words most related to trust and least related to trust. For 10 of the 12 union sets, the size of the union set was 150% or less than the *minimum* union set. Nine of 12 sets had a union set size that was 50% or less than the *maximum* union set size. These percentages, as well as the union set size and maximum and minimum set sizes, are plotted in Figure 2, for the sets of words most negatively and positively related to trust. Thus, while the word sets are not identical across conditions, the relatively small set size compared to the maximum union set size indicates a reasonable degree of similarity across conditions. It should be noted that for the larger word sets, it is more likely that the sets will overlap. Since there were fewer than 90 words that were positively related to trust in the set participants were asked to rate, the sets of 30 words most related to trust had to overlap across the three conditions. However, the degree of overlap was similar across the small and large sets, indicating that the overlap was not due simply to set size, but rather to similarity in the meaning of trust across the three conditions.

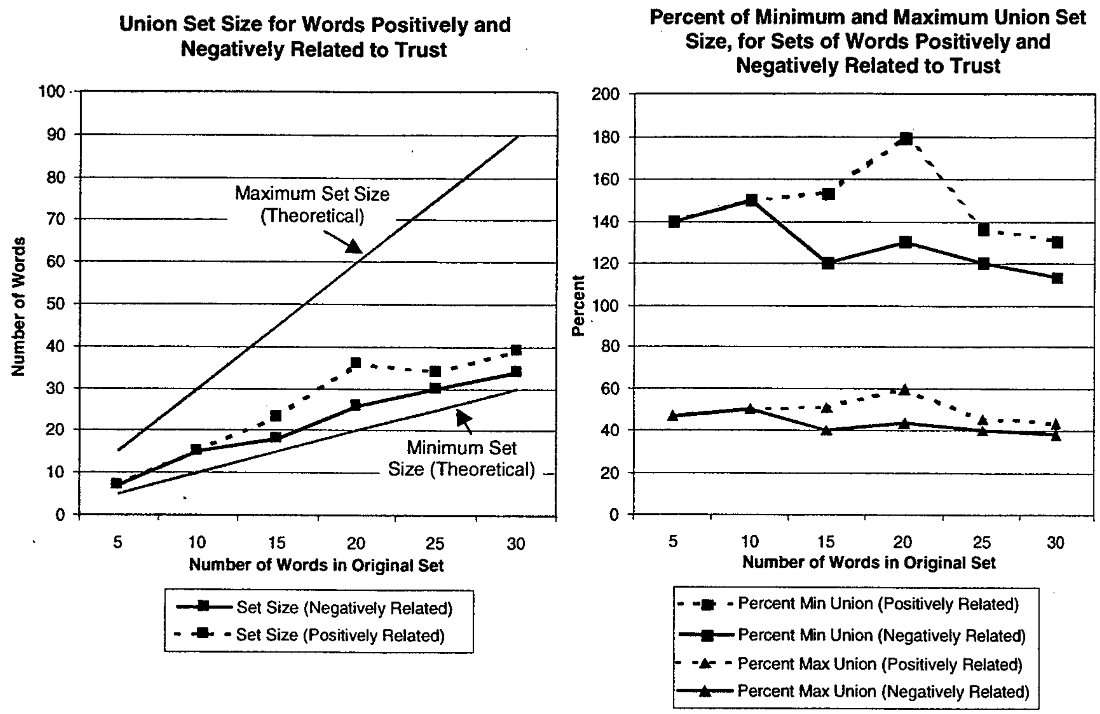

Figure 2. Plot of the union set size for the top 5, 10, 15, 20, 25, and 30 words most negatively and positively related to trust. The maximum and minimum union set sizes are provided for comparison, as well as the union set sizes' percent of the maximum and minimum set sizes.

Based on these results, words from the "top 10" set for each condition, positively and negatively related to trust, were selected to form the set of words for the next experimental phase, the paired-comparison study. These were 15 words in the "top 10" set negatively related to trust, and 15 words in the "top 10" set positively related to trust, for a total of 30 words. The final set of words, which we will refer to as Set-2, contained 30 trust and distrust related words. Set-2 was used in the subsequent computerized paired-comparison experiment. These words (Set-2) are shown in the three left-hand and three right-hand columns of Figure 3.

Less Similar to Trust | More Similar to Trust

Original Set Sizes	5	10	15	20	25	30	30	25	20	15	10	5
Trust Rating Range	-2.8…-1.8	-1.8…-1.5	-1.5…-1.3	-1.3…-1.1	-1.1…-.09	-0.9…-0.7	1.0…1.1	1.1…1.2	1.2…1.3	1.3…1.5	1.5…1.7	1.7…2.5
Size of Union	7	15	18	26	30	34	39	34	26	23	15	7
Percent of Min Union	140%	150%	120%	130%	120%	113%	130%	136%	130%	153%	150%	140%
Percent of Max Union	47%	50%	40%	43%	40%	38%	43%	45%	43%	51%	50%	47%
Words in Union Set from all 3 conditions (general trust, human-human trust, human-machine trust)	Betray Cheat Deception Distrust Mistrust Phony Steal	Betray Beware Cheat Cruel Deception Distrust Falsity Harm Lie Misleading Mistrust Phony Sneaky Steal Suspicion	Betray Beware Cheat Cruel Deception Distrust Falsity Harm Lie Misleading Mistrust Phony Selfish Skepticism Sneaky Steal Suspicion Wariness	Anger Attack Betray Beware Biased Cheat Cruel Deception Denial Distrust Doubt Error Falsity Harm Lie Misleading Mistrust Overcharge Phony Selfish Skepticism Sneaky Steal Suspicion Wariness Wrong	Anger Attack Betray Beware Biased Caution Cheat Cruel Deception Denial Distrust Doubt Error Failure Falsity Harm Hesitation Lie Mistake Misleading Mistrust Overcharge Phony Selfish Skepticism Sneaky Steal Suspicion Wariness Wrong	Ambiguity Anger Apprehen- sive Attack Betray Beware Biased Caution Cheat Cruel Deception Denial Dispute Distrust Doubt Error Failure Falsity Harm Hesitation Lie Loss Mistake Misleading Mistrust Overcharge Phony Selfish Skepticism Sneaky Steal Suspicion Wariness Wrong	Absolute Assurance Certainty Closeness Commit Competence Confidence Cooperation Credit Definite Entrust Faith Familiarity Fidelity Friendship Guardianship Honesty Honor Integrity Intimacy Love Loyalty Moral Pledge Positive Predictability Principle Promise Reliability Respect Responsibility Security Sincere Solid Stable Surety Trustworthy Understand- ability Unerring	Assurance Certainty Closeness Commit Competence Confidence Cooperation Definite Entrust Faith Familiarity Fidelity Friendship Guardianship Honesty Honor Integrity Intimacy Love Loyalty Moral Pledge Positive Predictability Promise Reliability Respect Responsibility Security Sincere Stable Surety Trustworthy Understand- ability	Assurance Certainty Confidence Entrust Faith Familiarity Fidelity Friendship Guardianship Honesty Honor Integrity Intimacy Love Loyalty Moral Pledge Positive Promise Reliability Respect Responsibility Security Sincere Trustworthy Understand- ability	Assurance Certainty Confidence Entrust Faith Familiarity Fidelity Friendship Guardianship Honesty Honor Integrity Love Loyalty Moral Promise Reliability Respect Responsibility Security Sincere Trustworthy Understand- ability	Assurance Confidence Entrust Familiarity Fidelity Friendship Honesty Honor Integrity Love Loyalty Promise Reliability Security Trustworthy	Familiarity Honesty Honor Integrity Loyalty Reliability Trustworthy

Figure 3. Union sets of the top 5, 10, and 15 words most negatively and positively related to trust. The size of the union, percent of the minimum and maximum union set sizes, and ranges of average ratings for the words in the set are also given.

EXPERIMENT THREE: PAIRED COMPARISON STUDY

The goal of the paired comparison study was to collect data for a subsequent factor analysis, in order to develop a multi-dimensional scale to measure trust.

Method

Participants

Thirty participants were recruited from members of the university community. All participants were native English speakers. There were 12 graduate students and 18 undergraduate students, of whom 14 were male and 16 were female. Participants were paid five dollars per hour for completing this one-session computerized experiment. Participants were told they could take a break at any time during the experiment and were required to have a short break every half hour. It took participants one to two hours to complete this experimental phase.

Procedure

Participants were asked to compare and rate the similarity of 30 words positively and negatively related to trust (a total of 435 pairwise comparisons). Participants used a computerized rating program to rate each pair of words on a seven-point scale with end points of "Totally different" and "Almost the same" (Zhang et al., 1996) by clicking on the appropriate rating (see Figure 4). A training session was conducted before the main program in order to familiarize participants with the task. Word pairs were randomized across participants.

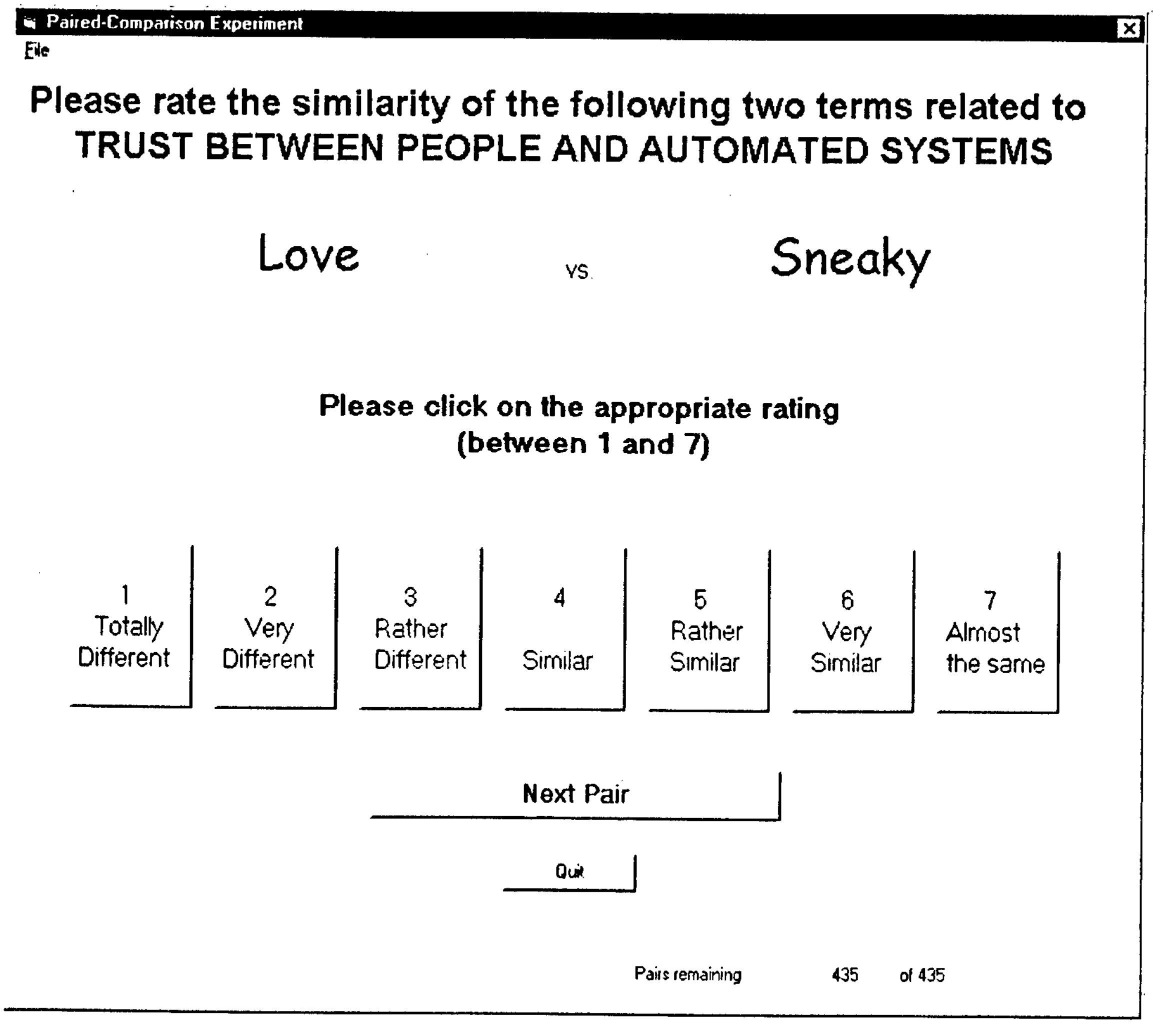

Figure 4. Example screen from the paired comparisons experiment.

Reliability Results

The similarity ratings from each participant formed a 30 by 30 similarity matrix. We performed an analysis on the similarity matrices for each condition to determine the reliability of the ratings, given the number of participants used. The sum of squares of index differences, *S(n),* was used in order to evaluate the stability of the structure. *S(n)* was defined as the sum of squares of index difference between the average similarity ratings of the first *n* participants and the average similarity ratings of the previous (*n*-1) participants:

$$S(n) = \sum_{i=2}^{30} \sum_{j=1}^{i-1} [A(n)_{ij} - A(n-1)_{ij}]^2$$

where $A(n)_{ij}$ and $A(n\text{-}1)_{ij}$ are the average similarity ratings of item *i* and *j* by the first n and (*n*-1) participants respectively. *S(n)* for each similarity matrix, as shown in Table 3, become small after eight or nine participants, indicating that the similarity matrix of ratings generated by 10 participants, as captured, can be considered reliable.

Table 3. The Reliability Values, *S(n)*, of Similarity Matrices for Three Conditions

Number of participants	1	2	3	4	5	6	7	8	9	10
General Trust	8247.0	309.0	54.8	63.3	60.7	22.2	12.7	6.8	2.5	3.2
Human-Human Trust	6851.0	277.3	59.8	12.4	6.2	10.6	10.4	17.7	6.4	6.5
Human-Machine Trust	2316.0	450.0	233.8	31.9	26.5	18.2	17.0	13.8	7.0	3.2

SCALE DEVELOPMENT

Two classification analyses, factor analysis and cluster analysis, were performed on data gathered in the previous phases in order to construct a multi-dimensional scale to measure trust.

Factor Analysis

The relatedness of words to trust or distrust obtained from the questionnaire study were analyzed by factor analysis using Minitab. Factor extraction using the principle components and varimax rotation resulted in nine significant factors for the condition of general trust, six in human-human trust, and eight in human-machine trust. Figure 5 shows the groupings of trust-related words for each factor. We determined the number of significant factors by selecting the top set of factors whose loadings explained at least 75% of the variance. The top set of factors for general, human-human, and human-machine trust explained 77%, 77%, and 79% of the variance, respectively.

Inspection of Table 4 shows that there are more groupings of positive trust-related concepts than negative ones. Additionally, there are fewer factors associated with human-human trust. From Table 4, we see that the smaller number of factors is due not to less differentiation in trust concepts (as would be indicated by fewer groupings), but rather due to the fact that more groups of related terms fell at opposite ends of the same factors.

Finally, we were able to identify some preliminary components of trust by examining these factors. First, the word, "familiarity" was extracted as a single factor across three conditions of trust. This indicates that people perceive *familiarity* as a unique component of trust, with respect to the other trust-related words. Second, the terms *assurance*, *confidence*, and *security* were grouped as a factor of both human-machine and general trust (*friendship* also appeared in the equivalent general trust factor). This factor may reflect a component of "confidence" in human-machine and general trust. Human-machine trust also had a factor combining *entrust*, *trustworthy*, and *reliability*, perhaps reflecting a component of "reliability" specific to human-machine trust. In contrast, in human-human trust, the concepts of *confidence* and *reliability* were grouped in a single factor. Separate factors of *familiarity*, *reliability*, and *confidence* are consistent with Sheridan's (1988) components of trust.

Table 4. Words Comprising Different Factors of Trust, for Three Conditions

Factor	Negative Grouping	Positive Grouping	Variance Explained
General Trust			
1	Cheat Betray Deception Steal Suspicion Distrust	Honesty Loyalty Love	0.178
2	Sneaky Misleading Mistrust Phone	n/a	0.116
3	n/a	Confidence Assurance Friendship Security	0.087
4	Beware	Integrity Fidelity	0.076
5	n/a	Familiarity	0.071
6	Harm Falsity	n/a	0.071
7	Lie	Honor	0.061
8	Cruel	Reliability	0.058
9	n/a	Trustworthy Entrust Promise	0.053
Human-Human Trust			
1	Mistrust Distrust Lie Misleading	Trustworthy Entrust Confidence Assurance Reliability Security	0.193
2	Harm Cruel	Familiarity Love Friendship	0.164
3	Falsity Sneaky Cheat Betray	Honesty	0.129
4	Suspicion Beware Deception	Honor	0.108
5	Steal Phony	Integrity	0.097
6	n/a	Fidelity Loyalty Promise	0.088

(Table 4 cont'd.)

Human-Machine Trust			
1	Betray Deception Sneaky Steal	Fidelity	0.143
2	Distrust	Promise Loyalty Love Honesty Friendship	0.141
3	Lie Mistrust Cheat Harm	Trustworthy Entrust Reliability	0.127
4	n/a	Security Assurance Confidence	0.090
5	Suspicion Falsity Cruel Beware	n/a	0.089
6	n/a	Integrity Honor	0.070
7	n/a	Familiarity	0.069
8	Phony Misleading	n/a	0.064

However, in general, the factors were difficult to interpret in terms of scales of trust. Recall that factor analysis groups words according to their inter-correlations with a defined concept, in this case, correlations between ratings of each word's similarity to trust. It could be the case that words are similarly related to trust, but not related to each other. Thus, we conducted a cluster analysis of the paired comparison data to attempt to group trust-related words according to their similarity to each other.

Cluster Analysis

Cluster analysis was used to group words according to their similarity to each other, as measured in the paired-comparison study. The between-group average linkage method was performed using SPSS. From the factor analysis reported earlier, between 11 and 13 "groups" of words were found for each type of trust[1]. We used these results to inform our choice of "cuts" in the cluster analysis trees, attempting to obtain a similar number of groupings. We first selected a level of similarity to cut the human-machine trust tree, since that is the type of trust of most interest to us. We then cut the other two trees at the same level of similarity. Figures 5, 6, and 7 show the cluster trees in three conditions of trust. The vertical line indicates the cutting point, and left parentheses indicate the resultant clusters of words. Table 5 shows words in each cluster across the three types of trust. At the most general level, two main clusters, relating to trust and distrust respectively, were formed for both groups across general trust, human-human trust, and human-machine trust.

In order to compare the similarity of ordering across the three conditions, a rank order correlation analysis was performed on the ordering of words across the three conditions. Results indicated a high similarity of ordering for the three types of trust: general trust and human-human trust had a rank correlation of $r = .84$, general trust and human-machine trust had a rank correlation of $r = .88$, and

---

[1] Recall that although there were between six and nine significant factors for each condition, some factors contained both positive and negative groupings of words.

human-human trust and human-machine trust had a rank correlation of $r = .89$. This result indicated that the ordering of words according to their rated similarity was similar across the three conditions.

Comparing across groups, we can identify several similarities and differences. For example, a category linking *cruel* and *harm* was found across the three groups, perhaps reflecting a category associated with an injurious outcome. *Falsity*, *lie*, and *deception* were also grouped together across the three conditions. *Beware* and *familiarity* formed separate clusters across the three groups. *Fidelity* formed a single cluster in human-human trust, but was paired with *loyalty* in the other two conditions. Additionally, the word *suspicion* seems to have some similarity to *mistrust* and *distrust*. It was grouped with *distrust* in general trust, and both *distrust* and *mistrust* in human-machine trust.

Based on the results of the cluster analysis, we developed a proposed trust scale for human-machine trust, which included 12 items for measuring trust between people and automated systems. The 12 items were derived by examining the words in the empirically derived clusters for human-machine trust. Table 5 shows the 12 items with respect to groupings of words, while Figure 8 shows how the proposed scale might be presented to participants.

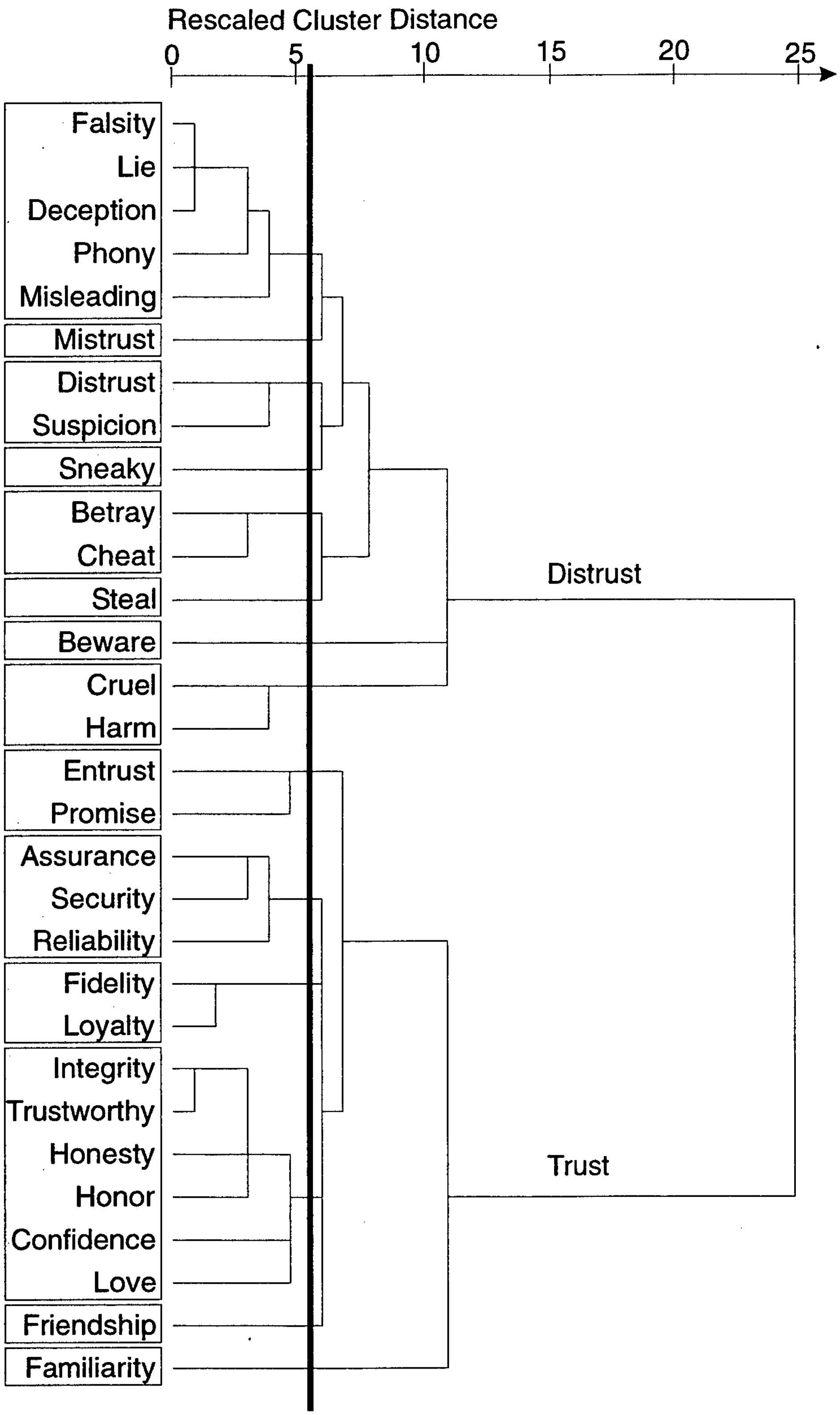

Figure 5. Cluster analysis for general trust. The vertical line shows the cutting point, and the shaded rectangles on the left-hand side of the figure show the resultant clusters. Notice the two large clusters corresponding to trust and distrust.

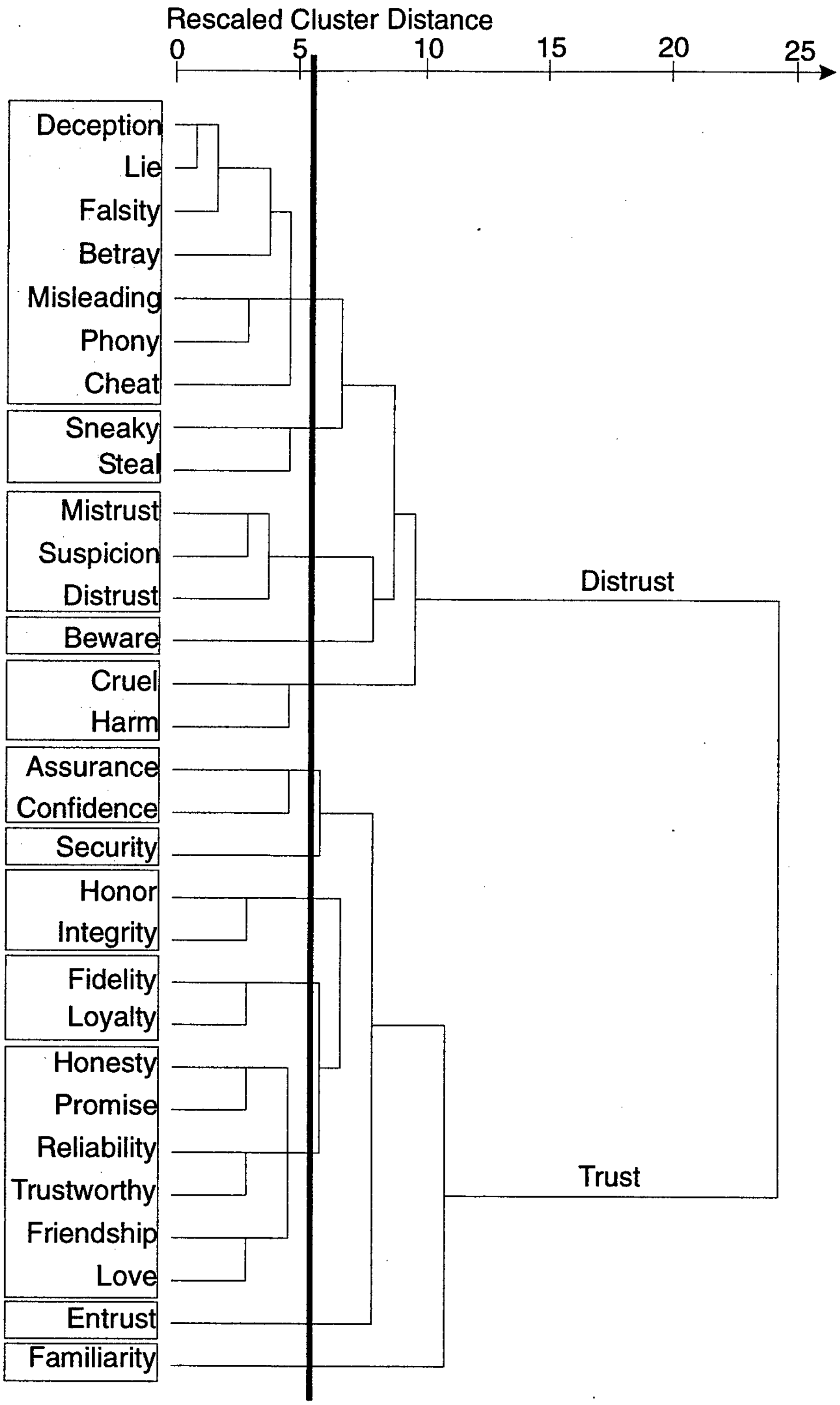

Figure 6. Cluster analysis for human-human trust. The vertical line shows the cutting point, and the shaded rectangles on the left-hand side of the figure show the resultant clusters. Notice the two large clusters corresponding to trust and distrust.

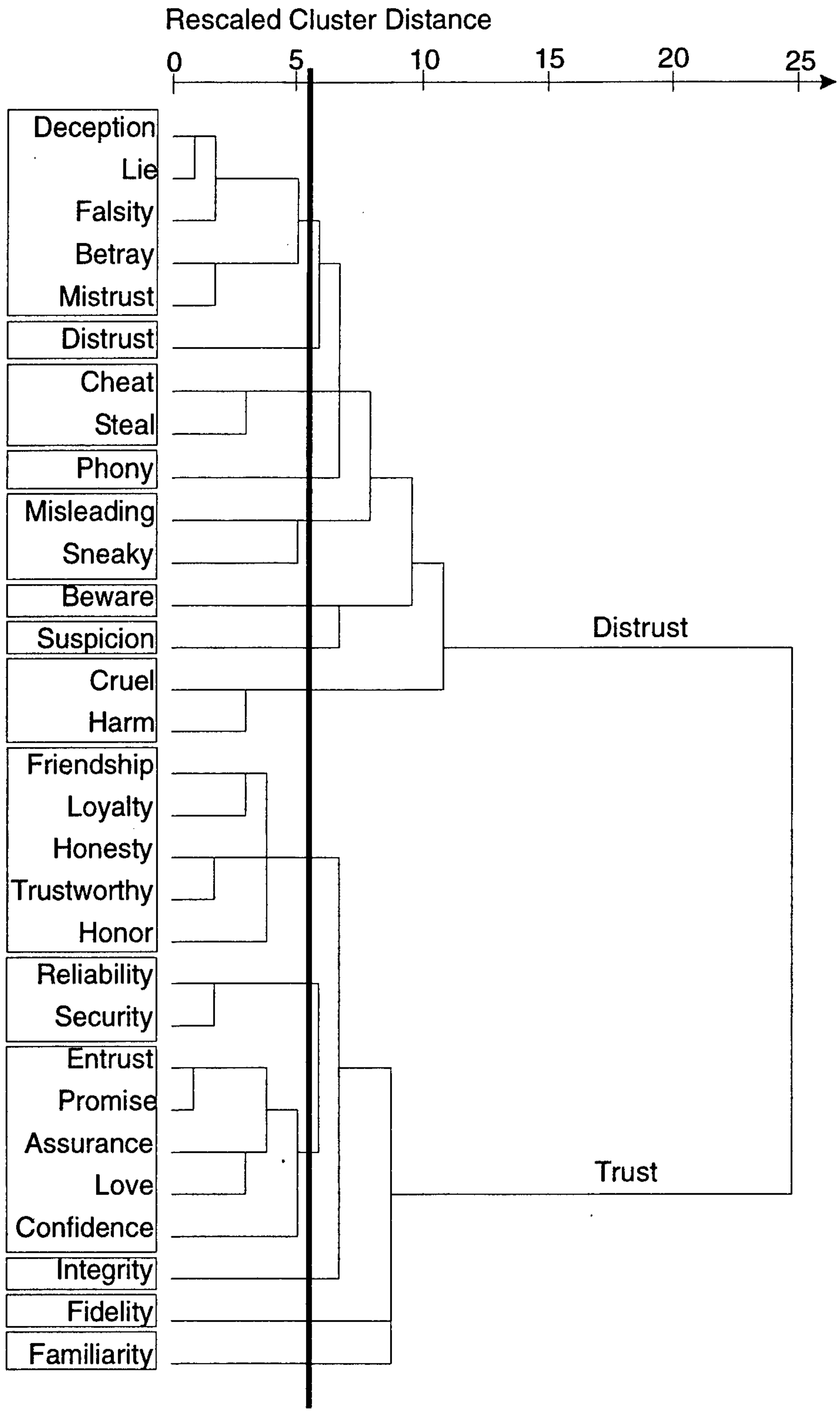

Figure 7. Cluster analysis for human-machine trust. The vertical line shows the cutting point, and the shaded rectangles on the left-hand side of the figure show the resultant clusters. Notice the two large clusters corresponding to trust and distrust.

Table 5. Trust Scale Items For Human-Machine Trust and the Corresponding Cluster of Trust Related Words on Which They Were Based

Item	Words Groups from Cluster Analysis
The system is deceptive	Deception Lie Falsity Betray Misleading Phony Cheat
The system behaves in an underhanded manner	Sneaky Steal
I am suspicious of the system's intent, action, or output	Mistrust Suspicion Distrust
I am wary of the system	Beware
The system's action will have a harmful or injurious outcome	Cruel Harm
I am confident in the system	Assurance Confidence
The system provides security	Security
The system has integrity	Honor Integrity
The system is dependable	Fidelity Loyalty
The system is reliable	Honesty Promise Reliability Trustworthy Friendship Love
I can trust the system	Entrust
I am familiar with the system	Familiarity

Checklist for Trust between People and Automation

Below is a list of statement for evaluating trust between people and automation. There are several scales for you to rate intensity of your feeling of trust, or your impression of the system while operating a machine. Please mark an "x" on each line at the point which best describes your feeling or your impression.

(Note: not at all=1; extremely=7)

1 The system is deceptive

1 2 3 4 5 6 7

2 The system behaves in an underhanded manner

1 2 3 4 5 6 7

3 I am suspicious of the system's intent, action, or outputs

1 2 3 4 5 6 7

4 I am wary of the system

1 2 3 4 5 6 7

5 The system's actions will have a harmful or injurious outcome

1 2 3 4 5 6 7

6 I am confident in the system

1 2 3 4 5 6 7

7 The system provides security

1 2 3 4 5 6 7

8 The system has integrity

1 2 3 4 5 6 7

9 The system is dependable

1 2 3 4 5 6 7

10 The system is reliable

1 2 3 4 5 6 7

11 I can trust the system

1 2 3 4 5 6 7

12 I am familiar with the system

1 2 3 4 5 6 7

Figure 8. Proposed questionnaire to measure trust between people and automated systems.

DISCUSSION

The above experiments have provided results which are important to the development of an empirically developed measure of trust. First, the high negative correlations of ratings of trust and distrust indicate that these concepts can be treated as opposites, lying along a single dimension of trust. In previous studies, this has been assumed, but not empirically tested. In practical terms, this implies that it is not necessary to develop questionnaires to measure high and low levels of distrust, separately from high and low levels of trust. This greatly simplifies scale design.

Second, from the questionnaire study and cluster analysis, patterns of ratings were similar across three types of trust: general trust, human-human trust, and human-machine trust, as indicated by the high degree of similarity in sets of words related to trust. This implies that people do not perceive concepts of trust differently across the different types of relationships. Note that both the questionnaire and paired comparison studies were between-groups designs, so that similarities between word patterns are not an artifact of carry-over between conditions. Although there were some differences, the overall similarity indicates that future work on the development of trust measures might not have to treat these types of trust differently, and also that results from studies of human-human trust (e.g., those that examine stages in the development of trust; Rempel et. al, 1985) may indeed have applicability to situations of trust between humans and automated systems. This transfer of trust concepts from the sociological to human-machine domain had not previously been tested empirically.

Third, the proposed scale of trust between humans and automated systems provides a model for assessing trust between humans and machines based on empirical data. From a practical perspective, this scale has the potential to help understand how system characteristics might affect operators' perception of trust. Once validated, the proposed scale may also be useful in predicting joint human-system performance, by providing a simple measure of trust in the system.

In particular, the scale was developed with respect to a non-directed feeling of trust in automated systems, rather than trust in a specific system which the participants had experienced. In this way, the scale developed here is dissimilar from certain of those used in the social sciences (e.g., Larzelere and Huston, 1980) which asked participants about trust in their romantic partner. However, the scale was not developed to measure a general personality trait of being trusting, but was focused on trust in a specific type of system. A general propensity to trust automated systems could provide an anchor for the development of trust in a particular system under a particular set of circumstances, and thus a measurement of this general propensity could provide a baseline measure with which to predict trust in a particular system, and changes in that trust over time.

Results from these experiments will provide the basis for future work on trust scales. Specifically, the proposed trust scale should be validated in experiments designed to understand trust in automated systems. For example, participants' actions regarding the use of an automated control system or information source could be captured, as the quality of those systems changes. As the system performance or information source degrades, one would expect participants to rely on the system or information less, and also to rate the system lower on factors of some trust on the proposed trust scale. Such a corresponding change in process measures on the one hand, and rated measures of trust on the other, would provide validation of the proposed scale. Scale reliability can be investigated by comparing rated measures of trust components across different participants, or the same participants over time, to see if changes in the quality of system performance or information source had a consistent impact on participant ratings of trust using the proposed scale.

CONCLUSIONS

A three-phased experimental study of trust concepts was performed to develop an empirically based scale to measure trust in automated systems. The experiments explored similarities and differences in the concepts of trust and distrust, and among general trust, human-human trust, and human-machine trust. Results provided empirical evidence for considering trust and distrust to be opposites, suggesting that two scales do not need to be developed to measure trust and distrust separately. Additionally, concepts of general trust, human-human trust, and human-machine trust tended to be similar, although people seemed to consider human-human trust more in terms of trust than distrust. Finally, results from the cluster analysis were used to construct a proposed scale to measure trust in human-machine systems.

REFERENCES

Larzelere, R. E., & Huston, T. L. (1980). The dyadic trust scale: Toward understanding interpersonal trust in close relationships. *Journal of Marriage and the Family*, 595-604.

Lee, J. D. & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. *International Journal of Human-Computer Studies*, 40, 153-184.

Lerch, F. J., & Prietula, M. J. (1989). How do we trust machine advice? In Salvendy, G. & Smith, M.J. (Eds.), *Designing and using human-computer interface and knowledge based systems* (pp. 410-419). North-Holland: Elsevier Science Publishers.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. International Journal of Man-Machine Studies, 27, 527-539.

Muir, B. M., & Moray, N. (1996). Trust in automation: Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics*, 39(3), 429-460.

National Research Council (1997). Automation. In Wickens, C. D., Mavor, A. S., & McGee, J. P. (Eds.), *Flight to the Future: Human Factors in Air Traffic Control.* (pp.271-289). National Academy Press, Washington D.C.: Author.

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in close relationships. *Journal of Personality and Social Psychology*, *49*(1), 95-112.

Sheridan, T. B. (1988). Trustworthiness of command and control systems. *IFAC Man-Machine Systems*, 427-431.

Singh, I. L., Molloy, R., & Parasuraman, R. (1993). Automation-induced "complacency": Development of the Complacency-Potential Rating Scale. *The International Journal of Aviation Psychology*, *3*(2), 111-122.

Zhang, L., Helander, M. G., & Drury, C. G. (1996). Identifying Factors of Comfort and Discomfort in Sitting. *Human Factors*, *38*(3), 377-389.