CIFE CENTER FOR INTEGRATED FACILITY ENGINEERING

Accounting for Rater Credibility when Evaluating Construction Industry Service Providers

By

Martin Ekstrom

CIFE Technical Report #148
FEBRUARY 2004

STANFORD UNIVERSITY

If you would like to contact the authors, please write to:

c/o CIFE, Civil and Environmental Engineering Dept.,
Stanford University
Terman Engineering Center
Mail Code: 4020
Stanford, CA 94305-4020

STANFORD UNIVERSITY

IT
BYGG
OCH
FASTIGHET
2002

IT Construction & Real Estate 2002

ACCOUNTING FOR RATER CREDIBILITY WHEN EVALUATING CONSTRUCTION INDUSTRY SERVICE PROVIDERS

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF CIVIL AND ENVIRONMENTAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Martin Ekström

June 2002

ABSTRACT

This study investigates how source credibility theory can support reputation mechanisms in AEC electronic commerce. Researchers and commercial interests have developed rating mechanisms that support trust in primarily consumer-to-consumer electronic market places. In contrast to consumer electronic marketplaces, the raters in business-to-business communities are skilled and connected, necessitating a reputation mechanism to account for the relationship between the user and the rater. Source credibility theory is an area of communication science that explicitly studies and formalizes trust between human actors. A rating system based on source credibility offers several advantages over existing models including tested frameworks for aggregating ratings from different sources and validated scales for measuring a source's (rater's) credibility. In addition, the weights of a rater's ratings depend on user preferences instead on rater behavior, which decreases the amount of data required to calibrate the model. I have divided the fundamental research question: *How can source credibility theory support rating systems in the procurement of AEC services?*, into the two dimensions: operationalization and added value. To investigate the research question, I operationalized source credibility into a credibility-weighted rating model, which assigns weights based on rater credibility. Furthermore, in two experiments, a set of industry users applied a credibility-weighted tool and an unweighted tool to evaluate bids from AEC subcontractors. Both experiments showed with statistical significance that the credibility-weighted models predicted rater weights better than an unweighted model. This study therefore contributes a methodology to operationalize source credibility theory to calculate rater weights for AEC. The experiments also showed that industry practitioners varied their evaluations more, and also were more confident in their judgments, when using a credibility-weighted tool than when using an unweighted tool. This study therefore provides evidence that a credibility weighted rating tool adds value in the process of evaluating AEC subcontractors by increasing the decision-maker's confidence in the accuracy of the information provided by the rating tool. I claim that these findings have power and generality and contribute to the literature of AEC electronic commerce, AEC Bidding, reputation mechanisms in electronic commerce, and applicability of source credibility theory.

ACKNOWLEDGEMENTS

I would like to express my gratitude to all those who gave me the possibility to complete this thesis.

At Stanford I have enjoyed cooperating with several very qualified persons. I am deeply indebted to my supervisor Prof. Hans Björnsson whose help, stimulating suggestions and encouragement helped me in all the time of research for and writing of this thesis.

I am also obliged to my other committee members Professor Clifford Nass, Doctor John Kunz, Professor Raymond Levitt, and Professor Pamela Hinds. I also want to Teddie Guenzer and Leyla Unerdem at the Center for Integrated Facilities Engineering (CIFE) for their help with numerous administrative issues. I would also like to express my thanks to my fantastic office mates Andrew Arnold, Jung Ung Mind and Francis Ng, who all have provided great help and who also shown extreme patience towards my disrespect of order. CIFE students who have provided me with insights and advice include Ragip Akbas, Roxanne Zolin, Sheryl Staub, John Haymaker, John Taylor, Bonsang Koo, Ashwin Mahlingam and Monique Lambert. Outside CIFE, I have received very useful help from the statisticians Philip Beineke and Dylan Small who both helped me validate and develop the rating model. I am also very grateful to my writing tutors Mary McDevitt and Carolyn Sale who looked closely at many different versions of the thesis for English style and grammar, correcting both and offering suggestions for improvement.

This work is possible thanks to the generous funding from several sources. First, I would like to thank the CIFE members who funded the initial seed project. I am also indebted to the Swedish research organization FoU Väst (in particular Per Åhman and Ingvar Olofsson) for their generous support along with access to member companies. I also would like to extend my thanks to the CMDG group (in particular Andrew Arnold and Durwood Snead) for providing me with the possibility to work on this project during a summer. I have also enjoyed working on projects funded by Nihon-Unisys. Finally, I am grateful to the Sweden America Foundation for supporting my second year at Stanford.

Cooperation with industry has been instrumental to the success of this project. In particular, I would like to thank all the industry practioners at Skanska, Peab, Webcor, Swinerton, and DPR who have spent valuable time participating in my experiments.

Finally, this work is dependent on the support of friends and family. I have many great friends and cannot mention them all but Fredrik has not only been a great roommate but also shown great understanding towards my financial situation. Yimei, my girlfriend, has miraculously been able to stand me throughout the last two years, even though I have dedicated too much time and effort to my research rather than to more important things. Finally, I am obliged to my parents and my brother for their love, support and encouragement, even though I know they wished that Stanford were not located on the other side of the world.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Summary

This summary chapter follows the order in which the chapters are presented in the thesis. A summary of the theoretical departure precedes an outline of the research methodology. The executive summary ends with an integrated review of the Results and Contributions chapters.

1.1 Problem and Research Question

Rating mechanisms have been successfully deployed in consumer-to-consumer (C2C) electronic commerce. In contrast to the raters in consumer electronic marketplaces, the raters in business-to-business communities are skilled and connected, necessitating a reputation mechanism that will account for the relationship between the user and the rater. The user, who is typically a project manager or estimator at a large general contractor, can access ratings from known as well as unknown raters. In the user's opinion, ratings provided by an experienced friend will be more important those provided by an unknown project manager. The question is: How much more important? What weight does the user attribute to the ratings from each rater? The task of aggregating ratings from multiple sources becomes a key problem, as it is neither 1) straightforward to automate the task 2) nor feasible to solve this problem manually in a realistic industry setting.

Researchers as well as commercial interests have developed rating mechanisms for electronic market places. However, I identify three limitations of the applicability of existing rating mechanisms to bidding for services in AEC: 1) reliance on input parameters that were difficult to measure; 2) reliance on ad hoc operators; and 3) reliance on large datasets of rating/transaction data for calibration.

Source credibility theory is an area of communication science that explicitly studies and formalizes trust between human actors. My intuition was that a rating system based on source credibility had the potential to mitigate all of the problems identified above. First, source credibility theory provides validated frameworks for aggregating ratings from different sources. As a result, a rating system based on source credibility will avoid the use of ad-hoc operators to aggregate information.

Second, there are validated scales for measuring a source's (rater's) credibility; these can serve as the key input parameter in a rating system based on source credibility. Finally, the weights in a rating based on source credibility theory depend on user preferences and not on rater behavior, which decreases the amount of data required to calibrate the rating application. The opportunity to measure the credibility of the rater's organization as well as of the rating person further decreases the amount of user input needed.

The fundamental research question of this project is:

How can source credibility theory support rating systems in the procurement of AEC services?

I have divided this question into two sub-questions:

1) *How is it possible to operationalize source credibility to support the calculation of weights that are based on rater identity in an AEC rating tool?*
2) *How can a rating system based on source credibility theory add value in the process of evaluating AEC subcontractors?*

There are three important limitations in the scope of this research project. Firstly, the purpose is not to provide incentives for raters to be honest, but rather to help the user distinguish between reliable and non-reliable raters. Such a distinction helps the decision-maker obtain better information about supplier performance. Secondly, this research project focuses on the application of source credibility theory to aggregate ratings which are measured subjectively by peer industry practitioners. The scope of this investigation does not cover other types of information which can support bidding decisions, such as credit ratings, and project experience. Finally, from a transaction cost theory perspective [Williamson, 1991 #10], a rating system is likely to have the highest impact on transactions with a medium degree of asset specificity (investments designated to a specific transaction) for which the governance structure is typically hybrid. It is therefore appropriate to focus this investigation on subcontracting of services by AEC general contractors. These transactions involve medium/low to medium/high degrees of asset specificity and are generally governed using a hybrid model.

1.2 Point of Departure

1.2.1 Practical Point of Departure

The criteria that industry practioners use to evaluate subcontractor performance can be divided into three categories 1) Objective measurements (e.g., project experience), 2) Subjective measurements provided by a reputable third party (e.g., credit ratings), and 3) Subjective measurements provided by peer industry practitioners. The last category is the focus of this investigation, since it necessitates a rating system to consider the source of the rating. It comprises important performance measures such as maintenance of schedule, collaboration, quality of work, administrative skills, change orders, and payment of second tier subcontractors/ suppliers. In the current practice, this type of information is exchanged between industry practioners, who use gossip, interviews, reference checking, and, in some cases, internal rating systems. The problems with the existing methodologies are that they are very time consuming, and that information risks being lost or distorted.

In electronic commerce targeting consumers, rating applications such as those of eBay and Amazon.com have contributed to user adoption. In the AEC industry, BuildPoint and RatingSource provide rating applications to owners and contractors, while some large contractors such as NCC of Sweden have developed internal rating applications. None of the online rating applications which exist in the AEC industry considers the source of the ratings when calculating overall ratings. Outside AEC, Open Ratings, and epinions.com do base the weight of a rating on the identity of the rater, but it is not clear how to implement these systems in AEC. As shown, in the theoretical point of departure chapter, Open Ratings' collaborative filtering solution requires a large amount of data to function, while it is not clear that epinions' "Network of Trust" approach is consistent with the rationale of AEC practitioners.

1.2.2 Theoretical Point of Departure

The theoretical point of departure of this project covers research in four different fields: AEC electronic commerce, AEC bidding, Rating Mechanisms in Electronic Commerce, and Source Credibility Theory.

Few studies in the emergent field of research studying AEC electronic commerce have focused on rating applications. Therefore, little research has investigated the applicability and added value of rating mechanisms. However, an AEC rating mechanism can formalize the third party information which Zolin et al [1] have identified as an important input in model of trust in virtual project teams. It could also support Tseng and Lin's [2] subcontractor procurement model.

In AEC bidding, researchers [3-5] have identified subjective information provided by peer contractors as important determinants of bid decisions. However, the importance of the sources of this type of information has generally been neglected in earlier research. Researchers [2, 6-11] have proposed models to support the evaluation of AEC subcontractors and contractors, but they have designed these models for the use within one company only, under the assumption that all the company's raters are equally knowledgeable and trustworthy. Furthermore, little research has studied the added value of rating mechanisms in AEC bidding.

Outside construction engineering and management, there is an emergent field of research focusing on rating mechanisms in electronic commerce. Several researchers [12-14] have investigated the added value of rating mechanisms in C2C electronic commerce. Investigations show, for example, that eBay sellers with higher ratings benefit from higher prices [12-14] and increased probability of selling their goods[15]. Ratnasingham and Kumar [16] argue that in B2B electronic commerce it is also important to consider the existing relationships between the human actors involved in the transactions. Taking a more applied approach, several researchers have proposed alternative bases for rating applications such as collaborative filtering [17], reputation hierarchies [18], statistical filters [19], fuzzy logic [20], network of trust [21], and rules based mechanism [22]. I argue that none of these solutions can satisfactorily deal with the problem of aggregating information in an AEC electronic market place since they either 1) rely on ad hoc aggregating functions (rule based mechanisms and fuzzy logic), 2) run into difficulty measuring the required input parameters (network of trust), or 3) require a substantial amount of data for calibration (collaborative filtering, statistical filters, and reputation hierarchies).

In communication research, source credibility theory relates to "the attitude toward a source of communication held at a given time by a receiver [23]." Researchers agree that source' credibility is multi-dimensional and the most common practice is to distinguish between a source's perceived expertise and trustworthiness [24]. The higher the trustworthiness and expertise the source is perceived to have, the higher the weight attributed to information coming from that source. Source credibility has been shown to be applicable in commercial settings [25-28] as well as for the judging of web content [29-31], but little research has investigated its applicability in electronic commerce. A rating system based on source credibility has the potential of overcoming all three of the problems identified above with other rating mechanisms. First of all, source credibility provides tested frameworks [25, 32, 33] for aggregating ratings from different sources. Researchers [34-37] have also developed and validated scales for measuring a source's (rater's) credibility which would be the key input parameter in a rating system based on source credibility. Finally, the weights in a rating based on source credibility theory depend on user preferences, rather than behavior, decreasing the amount of data required to calibrate the rating application. Researchers have also demonstrated the impact of factors, other than credibility, on the aggregation of information from multiple sources. A rating tool based on source credibility could further improve its performance by taking into account, for example, number of sources [38], feedback discrepancy [39], message framing [28], impact of organizational belonging [40], and time [26]. Based on the above discussion, I conclude that there exists an opportunity to research the extent to which source credibility theory can support rating applications in AEC electronic bidding.

Previous research credibility has shown source credibility theory to be applicable in online (e.g.,[31, 41]) as well as commercial settings (e.g.,[25, 42]). However, hitherto little research has investigated its applicability in commercial as well as online settings (i.e., electronic commerce). Electronic bidding in the AEC industry is a commercial, online situation where there exist substantial benefits from information sharing alongside incentives for deceit.

1.3 Research Methodology

This section discusses the research methods I have used to investigate the fundamental research question: *How can source credibility theory support rating systems in the procurement of AEC services?* The major research methodologies have been modeling and experimentation. Based on research of the current practice and the theoretical point of departure, I operationalized source credibility into a rating tool named TrustBuilder. TrustBuilder exists in two versions (*TrustBuilder I*, and *TrustBuilder II*) which served dual purposes: 1) they investigated the feasibility of operationalizing source credibility, and 2) they supported experiments (*Experiment I* and *II*) that investigated the added value of source credibility based rating systems in AEC. Table 1 provides a chronological summary of the key research activities of this project.

Table 1 Description of key research activities in chronological order

Research Activities in Chronological Order	Description of Activity
Designed and built TrustBuilder I	Designed basic subcontractor rating model incorporating source credibility to calculate rater weights
Evaluated TrustBuilder I in Experiment I	Evaluated applicability of source credibility to support an AEC rating tool in an experiment with non-experts
Designed and built TrustBuilder II	Incorporated lessons learnt in Experiment I when designing refined source credibility based rating model
Evaluated TrustBuilder II in Experiment II	Repeated Experiment I in a setting where industry specialists used refined model to evaluate subcontractors based on actual ratings

Since TrustBuilder II was a refined version of TrustBuilder I, the two models had the same theoretical framework and shared similar implementation characteristics. First of all, both models measured rater credibility by letting the user assess potential raters with the McCroskey [35] source credibility scale. The two models also accounted for the relationship between the user and the rater by distinguishing between three separate situations: situations in which 1) the user knows the rater, 2) the rater is unknown, but the user knows the organization to which the rater belongs, and 3) both the rater and the organization are unknown to the user. Moreover, in order to transform the source credibility scale measures to weights, both models let the user make pair-wise comparisons by evaluating hypothetical subcontractors that have been rated by two disagreeing raters. TrustBuilder I then minimizes the errors of the pair-wise comparisons using logistic regression while TrustBuilder II deploys an exponential conversion function. Finally, both tools provide a user interface where the user can evaluate subcontractors based on overall ratings, which have been calculated by weighting the individual ratings by rater credibility.

TrustBuilder I and II were each tested in one experiment respectively (*Experiment I* and *Experiment II*). Both experiments were within-subject designs where the type of rating tool was the differentiating factor, and the user the primary unit of analysis. The participants evaluated the overall performance of a set of subcontractors bidding on a project using different rating tools. The objective was to compare the performance of a credibility weighted rating tool (TrustBuilder I or II) to that of a standard, unweighted tool. However, the initial Experiment I also tested user behavior when the participants had no ratings available. To investigate the operationalization part of the research question, the experiments measured the errors in the pair-wise comparisons to determine which model best predicted rater weights. Both experiments investigated the added value of source credibility in AEC bidding by measuring the variation of the user's evaluation of overall subcontractor performance when using the different rating tools (assuming that the more confident the user is in the overall ratings, the more likely she is to vary her evaluation of overall performance), as well as the confidence expressed in the evaluations. In addition to incorporating different versions of the TrustBuilder tool, the two experiments differed in terms of the type of participants and the underlying rating data. In Experiment I, all participants had construction management experience, but they were non-experts at evaluating subcontractors. The participants of Experiment II, on the other hand, were Bay Area professionals with extensive experience in evaluating AEC subcontractors. Finally, in Experiment I the rating data were hypothetical while in Experiment II the participants had rated actual subcontractor performance.

1.4 Summary of Results and Research Contributions

Both experiments showed with statistical significance that the credibility-weighted model was a more accurate predictor of rater weights than an average (unweighted) model. In the second experiment, I also showed all the factors of the credibility-weighted model to be statistically significant predictors of rater weights. This research project has therefore given evidence that:

1) The TrustBuilder methodology operationalizes source credibility theory to calculate rater weights.

Both experiments also showed with statistical significance that the participants varied their decisions more using a credibility-weighted tool than when using an unweighted tool. Another outcome of the two experiments was evidence that the use of the credibility weighted rating tool increases the users' reported confidence in the ratings. Finally, the second experiment showed that industry practitioners found the credibility-weighted tool to be more useful than the unweighted tool. As a result, this research project does give evidence that:

2) A credibility weighted rating tool adds value in the process of evaluating AEC subcontractor by increasing the decision-maker's confidence in the accuracy of the information provided by the rating tool.

Based on these findings I claim to make research contributions in the four following fields:

AEC electronic commerce – I claim that this research project provides evidence that rating tools can add value in AEC electronic bidding. Furthermore this research project provides evidence that weighting ratings based on source credibility can add value in a rating system supporting AEC e-bidding. This study also contributes to theory of AEC electronic commerce by providing evidence that experimentation can be used by researchers to investigate the applicability and added value of tools that support electronic commerce in AEC.

AEC Bidding –This research project contributes to the state of research in AEC bidding by providing a methodology for the integration of subjective information from multiple AEC practitioners of varying reliability. More specifically, this contribution consists of a methodology to formalize source credibility to calculate rater weights in AEC depending on the user's perception of rater credibility. This study also provides evidence that source credibility theory can add value in AEC bidding by increasing the user's confidence in the accuracy of the information. This evidence constitutes another important contribution to the field of AEC bidding.

Rating Mechanisms in Electronic Commerce - This research project claims contributions to the state of research in Rating Mechanisms in Electronic commerce. Firstly, the results provide evidence that it is possible to formalize source credibility to support rating mechanisms in electronic commerce. Secondly, this research project shows that a rating system incorporating source credibility theory can add value in B2B electronic commerce transactions relative to a standard, unweighted rating mechanism.

Applicability of Source Credibility Theory – This research project contributes to the state of research in the applicability of source credibility theory by providing evidence that source credibility can be applied to construct the weights given to information from different sources in an online commercial setting, where there are substantial benefits from online information sharing, as well as opportunities for deceit.

2 Problem and Research Question

This chapter begins by discussing the problem of selecting the best bidder in AEC electronic commerce, and, more specifically, the difficulty of designing a rating mechanism which can support trust in AEC electronic commerce. I then identify source credibility theory as a potential solution basis for a rating system in AEC electronic commerce and present the associated research question: *How can source credibility theory support rating systems in the procurement of AEC services?* I also show how we can consider this question in terms of the two dimensions: operationalization and added value. The chapter ends by stating the limitations of the scope of this research project.

2.1 Problem

This section starts with a discussion of the general practical problem that faces industry practitioners when they evaluate AEC subcontractor in order to select the best bid. I then discuss why trust is a prerequisite for the adoption of electronic commerce in AEC before introducing rating mechanisms as a potential enabler of trust. Rating mechanisms have been very successful in consumer-to-consumer (C2C) electronic commerce but the following section shows that there are some fundamental differences between C2C and Business-to-Business (B2B) electronic commerce. Due to the closer ties between the actors in the market and the increased difficulty of performing the ratings, a rating system in a B2B setting such as AEC should take into account the identity of the rater. As a result, the task of aggregating raters from multiple sources becomes a key engineering problem. As illustrated in the last part of this section, this problem is neither 1) straightforward to automate, 2) nor feasible to solve manually in a realistic industry setting.

2.1.1 Practical Problem - Select the best bidder

GC & Co, a hypothetical midsize California general contractor, is bidding on a $4M dollar job to complete a new communication center. This is a public job for the city, which means that the bid process is competitive, and that the “lowest responsible bidder" is awarded the job. The bid is due at 3 PM and the time is now half past two. Paving is one subcontract included in the bid package, and Chuck Numbers, the chief estimator at GC&Co, has estimated its cost to be approximately $300,000. Currently, GC&Co has one bid at $320,000 from BayPave. BayPave is a well-known subcontractor that has worked for GC&Co before. Suddenly a fax comes in with a bid of $240,000 for paving from "PaveUSA", another paving subcontractor. Chuck has never heard of PaveUSA and therefore not solicited a bid from them. All that he knows is that they fulfill the minimum requirements of being licensed and bonded. Chuck is now in a dilemma since he knows that if he does not use PaveUSA, someone else will. He also knows that the difference of $80,000 between PaveUSA's and BayPave's bids is large enough to decide which GC will be awarded the project. Chuck therefore decides to include the unknown subcontractor PaveUSA on his bid list. In the end, GC&Co wins the contract but PaveUSA does not perform and finally goes insolvent during the project. The result is what Chuck defines as "a huge mess." The extra costs of schedule delays and finding an alternative paver cut GC&CO's $200,000 profit in half. Is there a way that Chuck could have found out in advance that PaveUSA was a "non performing" subcontractor?

2.1.2 Trust is a prerequisite for AEC e-commerce

E-commerce, "the enabling technology that allows businesses to increase the accuracy and efficiency of business transaction processing", [43] can decrease the costs of business-to-business (B2B) transactions [44] [45] and potentially revolutionize business process in the large and fragmented Architecture Engineering Construction (AEC) industry. By subscribing to a B2Bmarket place, Chuck could have, at a low cost, solicited bids from a large set of subcontractors and thus enjoyed increased market transparency and also, ultimately, lower costs and, potentially, better supplier performance.

In April 2000 approximately $1 Billion was invested in around 200 AEC e-commerce start-ups [46]. The problem is that if AEC industry practitioners are to use electronic market places to find new market partners, they must be certain that they can trust the participating organizations. In reality, the number of transactions conducted in these new AEC e-market places turned out to be very low, and only a handful of them were still in business by March 2002.

Welty and Becerra Fernandez [47] argue that not only technology but also trust are important pre-requisites if companies are to benefit from business-to-business electronic commerce. Therefore, one possible cause of the slow adoption of e-commerce in AEC is the participants' lack of trust in e-market places. General Contractors, subcontractors and material suppliers are unlikely to do business with firms that are unknown to them. How can Chuck be certain that the subcontractors bidding on an electronic market place are not the "PaveUSAs" of the industry?

2.1.3 Internet Based Rating Systems enable trust in Consumer to Consumer (C2C) market places

The Internet can support development of trust since it, as a network, possesses excellent means for selecting and synthesizing information [48]. This capability manifests itself in rating applications by which the market participants share information about one another. In virtually every transaction, the 40 million members of eBay's online community make decisions about whether or not to trust an unknown seller or buyer. When buying and selling everything from sports memorabilia to used cars, eBay's users can take advantage of a rating system. Based on an investigation of the eBay rating system, Reznick et al [15] argue that the Internet provides a superior mechanism for distribution of the information which supports trust decisions. Rating systems are important features of major consumer-to-commerce (C2C) (e.g., eBay), as well as Business-to-consumer (B2C) (e.g., Amazon.com) market places. Consumers buying on eBay or Amazon can, with little effort and at no cost, take part of the quality reviews made by peer consumers. Drummond [49] has pointed out the need for information sharing in business-to-business market places as well. If Chuck had access to a set of ratings of PaveUSA's previous performance, it is very likely that he would have made a different decision. With better information, Chuck could have pursued either one of three strategies, each of which would have increased the chances of a satisfactory outcome. Chuck could have 1) chosen to use the well-reputed BayPave rather than PaveUSA, 2) decided to use BayPave but added a risk buffer in his bid to the owner, or 3) used BayPave without adding a risk buffer, but made sure that CalGC would proactively manage PaveUSA during the project to mitigate problems before they occurred. As a result, an interesting research area is the applicability of rating systems in AEC electronic commerce.

2.1.4 Differences between B2B and C2C Marketplaces

Online rating tools have so far been slower to catch on in B2B electronic markets such as AEC, than in C2C e-commerce. I argue that the reason for the slower adoption is that there are a number of fundamental differences between B2B and C2C market places:

Information available to support trust decisions – The feedback provided by the rating application is often the only information that is available for the buyers and sellers at eBay to judge each other's trustworthiness. Decision makers in B2B electronic market places require more information than just peer ratings to support their decisions. The current practice of evaluating subcontractors in the AEC industry integrates different types of information from a wide variety of sources, including information from bonding institutes, data about past projects from market data providers, and references from peer contractors. Today, little or none of this information is available online but this situation is likely to change in the future. In a C2C marketplace, the vendor rating and price may be the only determinants of a consumer's choice of vendor, while in a B2B context, there are many other factors to consider.

Transaction Value – The value of B2B transactions is usually higher than for a C2C transaction. This leads to increased opportunities for deceit as well as changes in buyer behavior. There have been several instances of people providing deceitful ratings to, for example, eBay. Nonetheless, the risk involved in trusting an unknown seller is smaller for a consumer buying a $10 baseball card than for an estimator procuring a $300,000 paving subcontract. List and Lucking-Riley [50] have also shown that users behave in a more rational manner when the transaction amount is high. As a result, it is a fair assumption that the user of an AEC rating tool would want to ensure the raters' trustworthiness.

Legal Issues – This project does not address legal issues, which become more important in B2B compared to C2C electronic commerce. Several AEC managers, to whom I presented the concept of an AEC rating system, objected that there is a risk that a subcontractor, who has received a low rating, would sue the rater or the provider of the rating system. Furthermore, in several European countries, such as Sweden and France, laws prevent the creation of databases that store information about individuals and organizations.

Market structure – The communities of C2C market places often comprise thousands or millions of anonymous users. As a result, buyer and seller transactions are rarely repeated [15]. In the construction industry, general contractors often do a substantial part of their business with a small number of recurring subcontractors [51]. The probability that a decision maker will know the person who is issuing the ratings is much higher than in a C2C market place. A rating system should therefore take into account whether the user knows the rater. Chuck will trust a close friend more than an anonymous project manager. Another complicating factor is the integration of ratings from within and outside the user's organization. Chuck would normally trust a project manager at GC&CO more than somebody working for another contractor.

Difficulty of evaluating performance – The goods and services purchased in business-to-business markets are in general more complex than consumer goods. It does not, for instance, take much expertise to judge the performance of a seller of a used tennis racket on eBay. In B2B transactions, on the other hand, rater experience and competence become important factors. For example, when evaluating an AEC subcontractor, Chuck would like to know if the rater of PaveUSA has experience employing pavers. In general, the rater's expertise is generally higher in B2B than in

C2C. As a result, B2B decision-makers, such as the users of an AEC rating tool, are interested in knowing the identity of the raters.

In this research project, I focus on the last two differences between C2C and B2B market places: market structure, and difficulty of evaluating performance. In a market where many of the raters are known to the user and where the task of evaluating performance is difficult, it is clear that, in the user's opinion, some raters carry more weight than others. Reznick et al [52] recognize that in rating systems that support C2C electronic commerce there is a "potential difficulty in aggregating and displaying feedback so that it is truly useful in influencing future decisions about who to trust." I argue that, due to the increased importance of the identity of the rater, this problem is likely to be even more acute in B2B electronic commerce. The next section illustrates this argument through a specific example in AEC bidding.

2.1.5 Aggregating information from multiple sources

The purpose of this section is to illustrate through a hypothetical example that aggregating ratings from multiple raters of AEC subcontractors is a problem which is neither 1) straightforward to automate 2) nor feasible to solve manually in a realistic industry setting. Table 2 presents two scenarios where Chuck is evaluating PaveUSA's *ability to maintain schedule* based on a set of ratings. For simplicity, the scale of the ratings is binary with the two values "Good" and "Poor." In Scenario I, there are only two raters who have evaluated PaveUSA on this criterion.

Table 2 Two scenarios of ratings of PaveUSA

Scenario I

Rater	**Rating of PaveUSA’s Ability to maintain schedule**
Jim Murray, Friend of Chuck Numbers and Project Manager at GC&Co with 20 years experience	“Good”
1 Unknown Project Manager working for an unknown General Contractor	“Poor”

Scenario II

Rater	**Rating of PaveUSA Ability to maintain schedule**
Jim Murray, Friend of Chuck Numbers and Project Manager at GC&Co with 20 years experience	“Good”
25 Unknown Project Manager working for unknown General Contractors	All 25 ratings: “Poor”

The first Rater is Jim Murray, a close friend of Chuck with 20 years of industry experience, who has worked as both a project manager and an estimator for GC&Co. Jim rated PaveUSA’s performance as “Good.” An unknown project manager, working for a contractor Chuck is not familiar with, provides the second rating, which indicates that PaveUSA’s ability to maintain schedule is “Poor.” In Scenario I, it is very likely that Chuck would trust Jim’s judgment of PaveUSA’s performance and still accept PaveUSA’s bid.

In Scenario II, there are 26 ratings of PaveUSA’s ability to maintain schedule. Jim still rates PaveUSA’s performance as “Good” but there are now 25 unknown project managers rating PaveUSA schedule performance as “Poor.” Chuck still trusts Jim’s judgment more than that of the unknown project manager, but now PaveUSA’s performance on Jim’s job appears to have been an exception. In Scenario II, Chuck is alerted of PaveUSA’s potential schedule problems, and therefore seriously considers hiring another subcontractor. In Chuck’s opinion, Jim’s ratings should weigh more than the ratings from unknown project managers. The question is: How much more weight should they have?

In reality, the distribution of a set of ratings is likely to lie somewhere between the two extremes that the two scenarios present. Moreover, the two scenarios are limited to the ratings of one subcontractor on a single criterion, which makes it possible for the estimator to do the aggregation manually. However, in time critical competitive bid situations, the general contractor is often dealing with four to five bidders on twenty to thirty different trades. Let us assume that the rating system comprises five to ten rather than one criterion, and that, on average, five people have rated each subcontractor. The result is that the decision-makers at the general contractor would be dealing with 2000-7500 individual ratings. It would therefore not be feasible to manually study all the individual ratings available for each subcontractor. As a result, there is a need for models which aggregate information from different raters. The question is: *How can we aggregate ratings from raters of varying reliability in a manner that is consistent with decision-maker rationale?*

2.2 A rating system based on source credibility theory is a potential solution to the engineering problem

This section argues that the existing solutions are insufficient when it comes to aggregating ratings and presents my intuition that source credibility theory is an alternative solution to this problem.

2.2.1 Problems with existing solutions

The point of departure chapter of this thesis shows that there exists no rating application in current AEC practice which satisfactorily deals with the problem of aggregating ratings from multiple sources. Furthermore, my investigation of rating applications proposed by construction management and engineering researchers shows that 1) several of them recognize the importance of subjective information supplied by peer industry practitioners, but also that 2) none of the proposed subcontractor evaluation tools accounts for the identity of the source of the ratings.

Outside construction engineering and management, there is an emergent field of research which focuses on rating mechanisms in electronic commerce. Several researchers have proposed alternative bases for rating applications such as collaborative filtering [17], reputation hierarchies [18], statistical filters [19], fuzzy logic [20], network of trust [21], and rules based mechanism [22]. These solutions have significant limits since they 1) rely on ad hoc aggregating functions (rule based mechanisms and fuzzy logic), 2) run into difficulty measuring the required input parameters (network of trust), or 3) require a substantial amount of data for calibration (collaborative filtering, statistical filters, and reputation hierarchies).

2.2.2 Source Credibility Theory an alternative approach to calculate rater weights

Source credibility theory is field in communication research which relates to "the attitude toward a source of communication held at a given time by a receiver [23]." There is widespread agreement that a source's credibility is multi-dimensional, and the most common practice is to distinguish between a source' perceived expertise and trustworthiness [24]. We should also note that researchers have shown source credibility to be applicable in commercial settings [25-28] as well as for the judging of web content [29-31]. However, little research has investigated its applicability in electronic commerce. Above, I identified three problems associated with implementing existing rating applications in AEC. My intuition is that a rating system based on source credibility has the potential of overcoming all three of these problems. Firstly, source credibility provides tested frameworks [25, 32, 33] for aggregating ratings from different sources. Researchers [34-37] have also developed and validated scales for measuring a source' (rater's) credibility, which would be the key input parameter in a rating system based on source credibility. Finally, the weights in a rating tool based on source credibility theory would depend on user preferences instead of rater behavior, which decreases the amount of data required to calibrate the rating application. Researchers have also demonstrated the impact of factors, other than credibility, on the aggregation of information from multiple sources. As a result, a rating tool based on source credibility could further improve its performance by taking into account, for example, number of sources [38], feedback discrepancy [39], message framing [28], impact of organizational belonging [40], and time [26]. I conclude that there exists an opportunity to research the extent to which source credibility theory can support rating applications in AEC electronic bidding.

2.3 Research Question

The fundamental research question of this research project is:

How can source credibility theory support rating systems in the procurement of AEC services?

I have divided this question into two sub-questions: the first focusing on the development of a methodology to operationalize source credibility in a rating tool, and the second investigating the added value of such a tool

Sub Question 1) ***How is it possible to operationalize source credibility to support the calculation of weights that are based on rater identity in an AEC rating tool?***

A methodology that operationalizes source credibility should satisfactorily deal with the following three issues: i) input parameters, ii) conversion function to translate the input parameters to weights, and iii) methodology to register user preferences:

i.) *Input Parameters:* *What parameters are required to model rater credibility in AEC electronic bidding?* In an AEC e-market, there is a possibility for a wide range of people belonging to different organizations to supply ratings. Earlier research has proposed source credibility scales which have been designed to primarily evaluate either persons that the user knows personally, or public persons or organizations. In order for users to take advantage of the entire knowledge of an e-market place, a rating tool should integrate information provided by raters whom the user knows very well, as well as raters who are completely unknown. Measuring rater credibility in AEC therefore requires parameters which go beyond the traditional dimensions of source credibility. At the same time it is important to limit the number of parameters in order to keep the model small and to ensure that it only includes factors which significantly contribute to the estimation of rater credibility.

ii.) *Conversion Function:* *What function can convert measures of source credibility to weights in a manner consistent with user preferences*? It is necessary to convert the input measures into a measure that models the user's estimate of rater credibility. Such a function should satisfactorily perform the job of estimating the weight that industry practitioners assign to different raters. These weights should be consistent with user preferences over a realistic input range.

iii.) <u>*Methodology to register user preferences:*</u> *What is an efficient methodology to register user preferences in order to calibrate the conversion function in a rating tool which supports AEC - bidding?* Finally, the conversion function contains coefficients that are specific for each user. One user may find expertise to be much more important than trustworthiness, while another user estimates all raters to be more or less equally credible. As a result, a rating tool requires a methodology to estimate the user specific coefficients in the conversion function. Such a methodology should efficiently capture user preferences that can serve as a target function in the estimation of these coefficients.

<u>Sub Question 2:)</u> ***How can a rating system based on source credibility theory add value in the process of evaluating AEC subcontractors?***

More specifically, this research project investigates the added value of a rating system in AEC in terms of user behavior and attitudes:

i.) <u>User Behavior</u>: *In what ways does the type of rating tool influence user behavior when evaluating AEC subcontractors?* For a decision support tool, such as a rating system, to add value, it should influence the industry practitioners' decisions. This research project investigates whether the type of rating tool affects decisions during the evaluation of AEC subcontractors which are bidding for a job. The value added by a rating tool can be studied by measuring the extent to which a rating tool can influence the decision-maker when she determines 1) the overall subcontractor quality and 2) the bid contingency, or risk buffer, added to the bids. When evaluating the overall subcontractor quality the decision-maker aggregates information which the rating tool provides in order to determine an overall rating. Based on the evaluation of overall subcontractor quality, he or she can then decide to take action by, for example, adding a risk buffer to the bid, deciding to hire the contractor, or making recommendations regarding the management of the subcontractor. The question is whether the type of rating tool that the decision-maker is using influences this decision. Another decision of particular interest is the bid contingency added to a subcontractor's bid. The bid contingency reflects the participants' assessment of the risk buffer that should be added to the bid, as well as the extra cost of managing an under-performing subcontractor. Ultimately, if enough contingency is added to the lowest bid, the decision-maker may decide to hire a subcontractor other than the lowest bidder.

1) User Attitudes**:** *In what ways do user attitudes depend on the type of rating tool they are using to evaluate subcontractors?* Another way to investigate the added value of a rating tool in AEC is to measure the decision-makers' attitudes about the different rating tools that they are using. Two interesting measures are confidence and usefulness of the rating tools. Comparing the users' confidence and assessment of tool usefulness makes it possible to draw conclusions about the added value of different rating mechanisms in AEC electronic bidding.

2.4 Project Scope

The study of rating systems supporting electronic commerce transactions opens up a wide field of research questions and topics. At this stage, it is important to state three important limitations in the scope of this research project.

This research project does not try to provide incentives for raters to be honest but to help the user distinguish between reliable and non-reliable raters. The facilitation of this distinction will help the users obtain better information about supplier performance. This goal is consistent with Reznick and Zechhauser's [52] observation that one of the primary challenges that faces the field of reputation mechanisms is to "provide information that allows buyers to distinguish between and trustworthy and non-trustworthy.” The only difference is that this project investigates the reliability of the rater instead of the seller.

The second limitation refers to the type of information used to evaluate subcontractor performance. In the Point of Departure chapter, I identify a number of performance criteria, which I classify into three categories: 1) Objective measurements (e.g., project experience), 2) Subjective measurements provided by a reputable third party (e.g., credit ratings), and 3) Subjective measurements provided by peer industry practitioners. This research project is primarily focused on the third category, “Subjective measurements provided by peer industry practitioners.” It is for this type of criteria that the source of the ratings is important and the number of ratings to integrate is high.

Finally this research project concentrates on the subcontracting of services in the construction industry. As section 3.2.1 of the point of departure will show, the impact of a rating system is likely to be the greatest for transactions which involve a medium level of asset specificity, and for which the governance structure is hybrid or a mix of a free market and a hierarchy. [The services that a general contractor subcontractors on a typical construction project are transactions that fulfill these characteristics, and they are also the focus of this study.] As result, the procurement of commodity products with low asset specificity, such as lumber, or highly specialized services with high asset specificity (e.g., specialized design) are beyond the scope of this project.

3 Point of Departure

This chapter presents this research project's point of departure from both a practical and a theoretical perspective. The practical point of departure describes current methods for evaluating bidders in the AEC industry and discusses existing rating applications in AEC as well as other industries. The chapter then establishes a theoretical point of departure through a discussion of existing research in four different fields, before ending with a discussion of the current research opportunities.

3.1 The Practical Point of Departure

This section establishes a practical point of departure for the current study by discussing the current practice for evaluating AEC subcontractors, along with the existing commercial rating applications. It begins with a description of the activities I performed to research the current practice. I then provide an in-depth discussion of the criteria that industry practitioners take into account when evaluating subcontractors. As this discussion shows, there exists substantial information sharing between peer AEC practitioners, even though this sharing takes place in an informal and rather arbitrary manner. In particular, the section investigates the existing methodologies for information sharing for those criteria that are measured subjectively, and where peer industry practitioners provide the information. The reason for this focus is that a rating system based on source credibility theory would primarily enable the sharing and aggregating of this type of information.

The subsequent section presents the existing online rating systems within AEC and other industries. It describes the major rating applications and discusses why these are insufficient when it comes to aggregating information from multiple raters in AEC. The section ends with a summary discussion, which suggests the practical implications of this research project for the AEC industry.

3.1.1 Activities performed to establish the practical point of departure

In order to investigate the current practice of evaluating subcontractors in AEC, I performed the following activities:

Interviews – I interviewed fourteen AEC practitioners to determine what methods they were currently using to evaluate subcontractors, and to identify the key requirements of an Internet based rating system. The informants included two project managers at general contractors, three estimators at general contractors, one insurance agent, one owner of a small subcontracting business, one architect, and four managers at a construction market data provider.

Field studies – I also performed two on-site observations at the office of a general contractor. The observations took place during the final critical hours before a bid was submitted to the owner. The field studies allowed me to study a group of estimators while they faced the task of choosing subcontractors for each trade under extreme time pressure.

Attending demonstrations of existing rating applications – To document the current state of the art in terms of AEC rating applications, I interviewed two groups of designers of intra-company rating tools. The groups were working for Buildpoint (an Internet Software provider based in San Mateo, California) and NCC (a large general contractor in Sweden). Both groups demonstrated the prototypes for their latest rating tools, which were still under development.

Collection of documents – I also collected the paper-based data used by different general contractors when evaluating subcontractors. The documents included bids, prequalification forms, scopes of work, trade journals, and printouts from existing rating applications.

3.1.2 Criteria used to evaluate subcontractors in current practice

The current practice among US general contractors evaluating the potential bidders on a project is, according to one chief estimator, “vague.” To evaluate the quality of potential subcontractors, the estimator uses judgment in combination with information from a variety of sources to evaluate subcontractor performance according to a number of criteria.

Based on my interviews with fourteen AEC practitioners, I determined a list of the most important criteria used by US general contractors to evaluate bidding subcontractors. I have divided the criteria into three categories 1) Objective measurements (e.g., project experience), 2) Subjective measurements provided by a reputable third party (e.g., credit ratings), and 3) Subjective measurements provided by peer industry practitioners.

3.1.2.1.1 Objective Measures

Subcontractor performance can be measured objectively in a number of ways. Quantitative measures, such as the number of times an event occurs, enable us to define an unambiguous scale, which makes it less important who provides the measures/ratings. Anyone measuring “the number of prisons that CalGC built in California 1995-2000” would probably arrive at the same number. Another feature of these measures is that they are easy to double check, if there is any doubt about the accuracy of the information. This research project does not focus on the evaluation of this type of criteria, since their aggregation is not dependent on the source. Listed below are the major quantitative criteria used to evaluate subcontractors in AEC.

Project Experience

Private publications such as CMDG [53] (Construction Market Data Group) and McGraw-Hill [54] publish information about the bidders on public jobs. These lists include information about which subcontractor has been selected by general contractor for different projects. A general contractor can therefore find out whether, for instance, a subcontractor has been hired by a competitor on repeated occasions. This would be taken as a sign of the subcontractor’s quality and stability. Public data on completed projects provide another source of information. On the Department of Transportation’s web site [55], for example, there is a list of all completed projects. This list identifies the contractors and if and when they were paid. However, this information will only be available if the contractor has been hired directly by the department and therefore does not provide information about subcontractors. Project experience can be divided into total and specific project experience. Both types of data are readily available online today.

Total Project Experience refers to the total number of contracts that a company has performed and informs the user whether a firm has a track record and whether it has been on the market for some time.

Specific Project Experience takes into account the context in which the project took place. Context specific factors include:

Type of subcontract – In some cases, it is important whether the two projects are of the same type. A general contractor could, for example, value a subcontractor who has experience building prisons.

Size of the project – If the subcontractor is bidding on a $1M subcontract but has previously only done $50,000 jobs, a general contractor may doubt that it has the resources to successfully complete the job.

For whom – A general contractor will generally value the choices made by a reputable competitor over the choices made by an unknown contractor. Today general contractors are incessantly trying to find out which subcontractors their competitors hire.

Geography – Has the subcontractor worked in this area before?

Safety: Worker's compensation modifier

General contractors often collect information about the subcontractor's worker's compensation modifier, which reflects the insurance rate that the contractor has to pay to insure the cost of accidents by their workers. The compensation modifier is a direct function of the number of accidents that the contractor has had during the recent years [56]. My interviews with estimators show that some estimators use a low worker's compensation modifier as an indicator of contractor competence.

Payment of union fees – A general contractor often deem subcontractors' failure to pay union fees as a sign of failure. Trade Unions can supply contractors with information about missed payments.

3.1.2.1.2 Subjective measurements provided by a reputable third party

Criteria which involve judgment but which are provided by an independent third party include, ratings from credit and insurance organizations. The third parties do apply elements of subjective judgment when assessing these criteria, but the general assumption is that the measures are consistent for all evaluated contractors. This research project does not focus on this type of criteria since, although the source is likely to influence the weight of the information (e.g., compare credit ratings from *Dun & Bradstreet* and *Dan & Barry*), the number of different sources is likely to be very small and often equal to one. Aggregation is therefore not a major issue for this type of ratings. Listed below are the major criteria involving subjective judgment which are provided by independent third parties.

Bond rate – The bond system in itself assures that the subcontractor will complete the job [57], and the subcontractor's bond rate is therefore a useful indicator of its financial stability. Given that subcontractors are bonded to perform the job (if not they will in most cases be automatically eliminated), the general contractor will study the bond rates of the subcontractors. A high bond rate indicates that the bonding institute associates a high risk with a given subcontractor. According to one chief estimator, "A bond rate of 2.5%-3% instead of 0.5% is a good indicator of problems and financial instability."

Credit Rating – Dun & Bradstreet (D&B) is the most well known provider of company credit ratings. D&B provides, for example, a measure called "Supplier Risk Score" [58], which estimates the risk associated with the supplier on a scale of 1 (Low) to 9 (High). The score is calculated using a statistical model derived from D&B's data files and reflects "the likelihood of a firm ceasing business without paying all creditors in full, or reorganizing or obtaining relief from creditors under state/federal law over the next 12 months."[58] However, it is important to note that, even though D&B uses statistics to calculate the ratings, several of the input parameters (company history, for example) are measured subjectively by D&B's employees.

Liability Insurance – General contractors want to know the name of the insurer and limit of the subcontractor's liability insurance to make sure that subcontractors can take on a job of the size they are bidding on. A high limit indicates that the contractor is prepared to undertake large contracts.

3.1.2.1.3 Subjective measurements provided by industry practitioners

The focus of this research project is criteria where the performance can, in a practical setting, only be measured subjectively by peer contractors. For this type of criteria, the source of the information is important and there is likely to be a high number of different ratings to be aggregated into an overall rating. In the preliminary interviews, I asked estimators and project managers to evaluate subcontractors based on two hypothetical divergent ratings supplied by peer project managers. The results showed that their opinions about unknown sources varied considerably. Some of the interviewees regarded the rating provided by an "unknown project manager" as equally important as ratings from a "trusted friend", while others would "never trust information" if they did not know the person supplying it.

Below I will first present the major subjective measurements provided by peer AEC practitioners, and then discuss the current practice for sharing this type of information.

3.1.2.1.3.1 Major Criteria measured subjectively by peer industry practitioners

Listed below are the major subjective measurements supplied by peer industry practitioners. The list shows that several important performance indicators belong to this category.

Maintenance of Schedule – Did the subcontractor maintain schedule on its projects?

Quality of Work – Was the quality of the subcontractor's work satisfactory?

Administrative Skills – A subcontractor's ability to smoothly handle all the paperwork associated with its relationship to the general contractor is, in some cases, a competitive advantage.

Collaborativeness –Is the subcontractor collaborating with the general contractor in case of contingencies and unexpected problems?

Change Orders – Is the subcontractor what is commonly known as a "Change order artist?" A well-known strategy in the construction industry is to bid low and make up the money on change orders.

Payment of 2nd tier subcontractors/ suppliers – Is the subcontractor paying its subcontractors and suppliers on time?

3.1.2.1.3.2 Methods for acquiring subjective information in current practice

In current practice, industry practitioners use a number of different methods to take account of their peers' assessment of subcontractor performance. Below I list the most important methods.

Reference Checking - Project Managers at competing general contractors (GCs) call each other to check the capability of subcontractors[59]. GCs also often ask subcontractors to provide references regarding jobs they have completed and project managers for whom they have worked. The use of references shows that there exists, in current practice, information sharing between industry practitioners about subcontractor performance, and that this exchange takes place within as well as across organizational borders. For negotiated jobs, reference checking is standard practice, but it is often more difficult to thoroughly investigate the quality of subcontractors on competitively bid jobs. This is because reference checking can be very time consuming, making it less helpful in time pressured bid situations. In addition, inferior knowledge about subcontractors is often seen as a strategic disadvantage for general contractors working in a new area. An out-of-town general contractor will often receive higher quotes from subcontractors than will local general contractors which are well known to the subcontractors [60]. Another problem is that the subcontractors themselves provide the references producing a positive bias. A subcontractor is unlikely to list as a reference a project where its performance was unsatisfactory.

Gossip – Many interviewees stressed the importance of gossip in order to find out, for example, "who's still in business", or "who's in trouble on what project." Gossip, or informal information sharing, between estimators also takes place at industry meetings such as AGC[1] meetings. These meetings act as a forum where the estimators can share information about the performance of subcontractors on an informal basis. It is also common practice for subcontractors to share information about the performance of general contractors. The subcontractors are primarily interested in knowing if the GC will pay on time, but also in finding out how good the GC is at managing projects. The problem with gossip is that a lot of information gets lost or distorted in the process.

Interviews – If time permits, the estimator or project manager will call the bidder to evaluate its expected performance [59]. The estimator wants to verify that the subcontractor "knows what he is bidding on" and that it has the technical competence necessary to successfully complete the job. Interviews are, of course, even more time consuming than reference checking and, as a consequence, even less feasible in time pressured bid situations.

Internal rating systems – Some larger general contractors document the performance of subcontractors in internal rating systems. One such system, which belongs to a large California contractor, classifies the subcontractors into five different categories:

1. *Preferred immediate Bay Area (typical Bay Area bid-list)*
2. *Preferred additional Northern California (typical Nor Cal bid-list-can include # 1s)*
3. *Reputable but not necessarily preferred*
4. *Not recommended*
5. *No information of performance*

By definition, the information in an internal rating system is shared only within the general contractor's own organization. Another feature of the above

[1] AGC: Associated General Contractors of America

described rating system is that it is uni-dimensional. It comprises only one overall criterion, which is based on the decision maker's subjective evaluation. Other California contractors have rating systems where three qualities, *technical competence, financial stability*, and *ability to cooperate,* are each evaluated subjectively.

Since these systems are mostly paper based, the process of entering and retrieving information to and from them is often very time consuming. Other problems include lost data and the considerable time it can take for a rating to become available to a user.

3.1.3 Existing Online Rating Systems

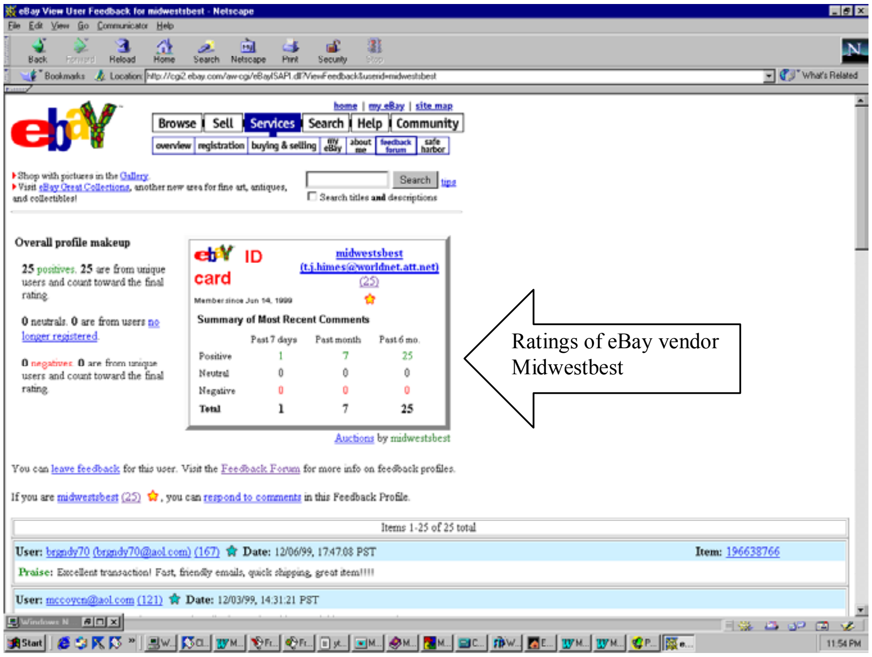
This section first describes the rating systems in electronic commerce outside AEC, before discussing existing the Internet based rating systems in the AEC industry.

3.1.3.1 Rating systems in electronic commerce outside AEC

E-commerce has had a faster penetration into the consumer market than in the B2B field. It is therefore natural that the earliest and most adopted commercial rating systems can be found in consumer-to-consumer e-commerce. Three of the ratings systems, which I describe below, belong to this area. The first, *eBay*, was created by a market maker to facilitate transactions between individuals. The second, *BizRate*, constitutes an independent third party information provider, which allows consumers to rate e-commerce vendors. The third, *epinions*, solicits ratings from consumers about products and services available off- as well as online. Finally, *Open Ratings* is a rating system, which was designed to support trust in B2B electronic market places.

3.1.3.1.1 EBay

EBay [61] is the largest and most successful Internet Auction, enabling transactions between private parties. As of March 2002, it has over 40 million registered users. Because items sold over eBay vary and are often difficult to describe, they cannot be easily evaluated. The buyer has to trust that the seller is describing the item fairly and the he or she will deliver the item once it is paid for. To foster trust, eBay has created a system by which the market participants rate each other after each transaction. The buyer will give the seller a good rating, if he or she received the items in good condition. The eBay rating system is simple and intuitive. After each transaction has taken place, the transaction parties rate each other on a three level scale ("positive", "neutral", and" negative") [62]. A user's total rating consists of the sum of all the ratings it has received during all the previous transactions. (+1 corresponds to a positive comment, 0 to a neutral, and –1 to negative.) Figure 1 below shows the rating for the user "Midwestbest," which is selling a Toulouse-Lautrec print on eBay. In this case, "Midwestbest" has a rating of 25 since it has received 25 positive comments but no neutral or negative comments. In addition to rating sellers and buyers on a numerical scale, the users are encouraged to submit comments. We find that user "brgndy70" who, in the past, has bought a "Maroon Panther Figurine" from Midwestbest, has added the comment: "Excellent transaction! Fast, friendly emails, quick shipping, great item!!!!" This information would make a potential buyer less reluctant to trust Midwestbest as a seller. EBay's success shows that it is possible to create trust over the Internet, even for goods where a substantial amount of trust is required.


eBay View User Feedback for midwestsbest - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Stop
Bookmarks Location: http://cgi2.ebay.com/aw-cgi/eBayISAPI.dll?ViewFeedback&userid=midwestsbest
What's Related
home | my eBay | site map
Browse Sell Services Search Help Community
overview registration buying & selling my eBay about me feedback forum safe harbor
Search tips
Start
11:54 PM

▸ Shop with pictures in the Gallery.
▸ Visit eBay Great Collections, another new area for fine art, antiques, and collectibles!

□ Search titles **and** descriptions

Overall profile makeup

25 positives. **25** are from unique users and count toward the final rating.

0 neutrals. **0** are from users no longer registered.

0 negatives. **0** are from unique users and count toward the final rating.

ID card

midwestsbest (t.j.himes@worldnet.att.net) (25)

Member since Jun 14, 1999

Summary of Most Recent Comments

	Past 7 days	Past month	Past 6 mo.
Positive	1	7	25
Neutral	0	0	0
Negative	0	0	0
Total	**1**	**7**	**25**

Auctions by midwestsbest

You can leave feedback for this user. Visit the Feedback Forum for more info on feedback profiles.

If you are midwestsbest (25) ★, you can respond to comments in this Feedback Profile.

Items 1-25 of 25 total

User: brgndy70 (brgndy70@aol.com) (167) ★ **Date:** 12/06/99, 17:47:08 PST **Item:** 196638766

Praise: Excellent transaction! Fast, friendly emails, quick shipping, great item!!!!

User: mccoycn@aol.com (121) ★ **Date:** 12/03/99, 14:31:21 PST

Ratings of eBay vendor Midwestbest

Figure 1 Screenshot showing ratings of a seller on eBay. During the past 6 months, the seller Midwestbest has received 25 ratings, all of which were positive.

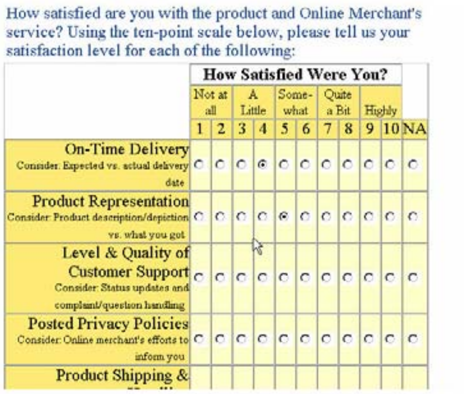
The problem with the eBay type of rating system is that that its ratings seem to be disproportionally positive; in fact, it is almost impossible to find a negative comment on eBay [15]. There are two possible major reasons for this. First, there are opportunities for collusion. Users can register under several user names and rate themselves and their friends positively. Another potential cause for the predominance of positive comments is that the sellers who have received negative ratings may scare off potential bidders. This in turn may create incentives for a seller to start from the beginning and re-register under a new identity, rather than continue with an old one.

The advantage of eBay's rating system is that it is simple and intuitive to use. A newly arrived customer has few difficulties interpreting what the different ratings mean. It is also quick and easy to rate someone after a transaction. As a result, the actual rating system adds only marginally to the transaction costs.

3.1.3.1.2 Bizrate.com

Bizrate is an independent site that allows consumers to rate different Internet businesses. Bizrate's revenues come from aggregate marketing research based on information supplied by the participating customers, and from the rebate checks they process. Bizrate rates only companies that have agreed to be rated. A business' participation in the rating mechanism can therefore function as a screening mechanism in itself.

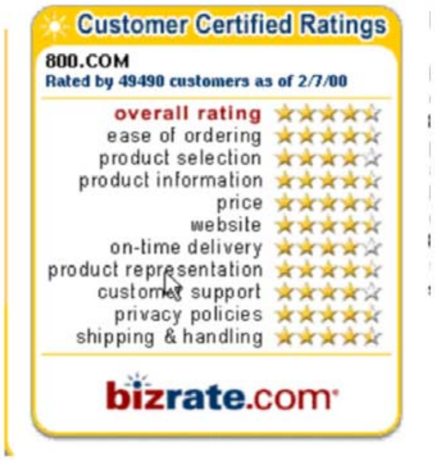
The Bizrate rating process is as follows. When a consumer has completed a transaction, he or she is asked to complete an initial survey.



How satisfied are you with the product and Online Merchant's service? Using the ten-point scale below, please tell us your satisfaction level for each of the following:

	How Satisfied Were You?										
	Not at all		A Little		Some-what		Quite a Bit		Highly		
	1	2	3	4	5	6	7	8	9	10	NA
On-Time Delivery Consider: Expected vs. actual delivery date	○	○	○	●	○	○	○	○	○	○	○
Product Representation Consider: Product description/depiction vs. what you got	○	○	○	○	●	○	○	○	○	○	○
Level & Quality of Customer Support Consider: Status updates and complaint/question handling	○	○	○	○	○	○	○	○	○	○	○
Posted Privacy Policies Consider: Online merchant's efforts to inform you	○	○	○	○	○	○	○	○	○	○	○
Product Shipping &											

Figure 2 a: Extract of Bizrate's follow up survey: evaluation the satisfaction relative to expectations



Customer Certified Ratings

800.COM
Rated by 49490 customers as of 2/7/00

overall rating	★★★★½
ease of ordering	★★★★½
product selection	★★★★☆
product information	★★★★½
price	★★★★½
website	★★★★½
on-time delivery	★★★★☆
product representation	★★★★½
customer support	★★★★☆
privacy policies	★★★★½
shipping & handling	★★★★½

bizrate.com

Figure 2 b: Bizrate's rating of the online merchant 800.com

Figure 2 Screen shots of Bizrate's rating application

Bizrate uses a ten-dimensional scale for its ratings. Some of the dimensions comprise criteria that can be evaluated before the goods are delivered. These include "Convenience and speed of ordering", and "Breadth and depth of products offered." Other criteria, such as "On-Time Delivery" and level of quality, cannot be evaluated until after the transaction. The customers are therefore asked to complete a second follow-up survey once they have received the goods. Bizrate also asks its customers to submit a survey indicating how important they perceive each of the ten dimensions to be. This information could potentially be used to refine the ratings, by weighting them according to their importance.

Bizrate presents the information about the participating business on a 1-10 star scale for each of its ten dimensions (see Figure 2 above). The ten dimensions are in turn aggregated to produce a single average value.

3.1.3.1.3 Epinions.com

Epinions.com is a site that functions like a community. At epinions.com consumers can exchange opinions about all types of products, from SUVs to kids' TV shows. The idea is that users who write reviews will earn recognition and rewards based on the usefulness of their advice, and "the focus of the service will be on qualitative reviews rather than quantitative rankings"[63]. To measure the usefulness of the advice, epinions allows the users of the advice to rate the advisors. A user can see how many other users trust the advisor, which enables her to quickly estimate the quality of the particular rater's advice. Another functionality is the "Web of Trust," which determines which opinions the rating system will display. The principle, as shown in

Figure 3, is to use "indirect trust" – when you trust someone because someone you trust trusts that person -- to find "trusted opinions". (The Section 3.2.4.2.1 *Network of Trust* section in the Theoretical Point of Departure discusses this approach in more detail.) The concept of rating raters can also be found in other community sites which provide advice, such as expertsite.com and Xpertcentral.com [64].

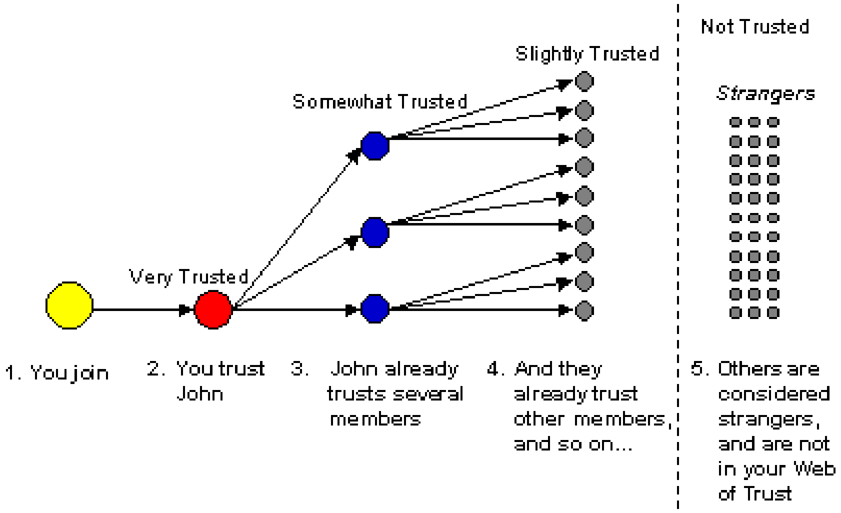

Not Trusted
Slightly Trusted
Strangers
Somewhat Trusted
Very Trusted
1. You join
2. You trust John
3. John already trusts several members
4. And they already trust other members, and so on...
5. Others are considered strangers, and are not in your Web of Trust

Figure 3 Epinion's "Web of Trust" model

Similar to these sites, Epinions enhances trust by encouraging its raters to provide personal information and build a reputation as a good rater. In interviews with AEC decision-makers, I have found that the attitude towards the idea of a network of trust varies considerably. For some interviewees, the concept of trusting "a friend of a friend" seemed intuitive, while others did not consider it to be relevant whether they and an unknown rater had a common friend.

3.1.3.1.4 Open Ratings

Open Ratings provides a rating service focused on supporting e-commerce market places, primarily in the B2B space. The original tool incorporated a complex weight algorithm developed by Zacharia [65] which combined features of Network of Trust, as well as Collaborative Filtering, both of which are described in section 3.2.4.2 of this chapter. Open Rating's algorithm for determining the weight of a rater takes into account the following parameters [66]:

- *The rater's previous ability to provide accurate ratings;*
- *The number of ratings completed by the rater;*
- *The length of time the rater has participated in the Open Ratings system;*
- *The rater's habits (this weighting mechanism prevents collusion and makes the ratings as meaningful as possible;)*

- *The size and circumstances of the transaction;*
- *The reputation of the ratee, based on previous rating; and*
- *Whether or not the rating is given anonymously.*

Due to the complexity of the rating filter, the actual aggregation is a black box from the users' perspective. The assumption is that the user will trust the overall rating that Open Ratings provides.

Recently, Open Ratings expanded its rating application by providing supplier risk analysis and monitoring tools. In the past, Open Ratings has tried to target the construction industry through partnerships with now defunct B2B e-commerce actors. These B2B e-commerce actors could have supplied the domain expertise, which adapting Open Ratings solutions to the construction industry would require.

3.1.3.2 Web based supplier evaluation systems in the AEC Industry

3.1.3.2.1 Ralacon and Godkjenningsordningene – Two Scandinavian prequalification systems

The Finnish AEC industry has taken a systematic approach to supplier screening through the creation of the prequalification system Ralacon. RALA, an organization whose members include the associations of Finnish contractors, architects and engineers, as well as owners and authorities, has created Ralacon. The owners have agreed to hire only contractors who fulfill the RALA requirements. The main objective is to exclude non-performing and semi-legal companies from the market. RALA issues certificates to those contractors who fulfill the RALA qualification requirements. Contractors demonstrate that they have fulfilled the Rala requirements by providing the following information:

1) General information about the company – Name, address, company ID, etc
2) Documentation from authorities: payment of taxes etc
3) Technical evaluation criteria
    i) Personnel
    ii) Reference work

4) Financial performance
5) Documentation of quality audits. ISO-9000 etc,
6) Areas of operation
    i) Specialties, types of works
    ii) Size of projects

A similar web-based search system for contractors exists in Norway [67]. The Norwegian government's list of qualified providers of AEC services currently comprises over 7000 companies. To qualify as a provider of a certain service (e.g., drilling of wells), a company must fulfill qualifications regarding its organization, management and skill sets. A company applying for qualification, submits documents such as an organizational plan, a quality management plan, proof of their administrative skills, manuals for completing the various services, a list of references projects, and information about its managers (their resumes, and certificates of education). If the government deems all these documents to be satisfactory, the contractor will then be listed in the Godkjenningskatalogen public database of qualified service providers for a period of 24 months.

To summarize, both Ralacon and Godkjenningsordningene are minimum standards supervised by an independent third party. A user will only know that a qualified member is above the minimum standard. There is no means to differentiate between qualified market participants, even though their quality may differ substantially.

3.1.3.2.2 EASY – an in-house rating system

The large Swedish contractor NCC has developed EASY (Evaluation Assessment System), an in-house supplier rating application [68]. The system supports two key business processes: *bedömning* (assessment of supplier performance prior to transaction) and *utvärdering* (evaluation of supplier performance after completion of the transaction). The suppliers/subcontractors are rated on ten criteria on a scale from one to three. All but two of the criteria are similar to the ones which American contractors find important such as schedule and collaboration (see section 3.1.2 for the complete list). However, NCC's criteria *Environmental performance* and *Quality of work environment* are less common in the U.S. When rating a supplier on the ten criteria, the raters also provide their motivation for each rating. The motivation will help later users to understand the context and rationale behind the rating. EASY employs a rule-based method to aggregate the ratings. A supplier is classified as a "class A supplier" if its average rating is higher than 70% and no criterion is rated is less then two. The system will warn the user if the supplier has received a low rating on a criterion, and provide the user with the motivation for the low rating. This way, managers can proactively manage suppliers to avoid potential problems before they occur. Finally, EASY's rule-based aggregation function weights all raters and criteria the same.

3.1.3.2.3 Eu-supply

Eu-supply provides online bidding solutions for the European construction industry. A bid process on eu-supply involves extensive preparation. The owner, with the aid of Eu-supply consultants, searches for, screens, qualifies, invites, and rates the performance potential bidders. During the bid phase, the owner can adjust the bids from the participating bidders based on "product, service, commercial terms and other company differences" [69]. If one bidder is deemed to be more risky than its competitors, its bids can be penalized with a risk buffer, or contingency, expressed as a percentage of the bid amount. The bidders do not know if the owner has adjusted either its rates or those of its competitors, and can only see how low it has to bid to beat the competing bids. The owner's preferences and risk profile determine the risk buffers added to the bids. Eu-supply's bidding application shows that the perception of supplier performance influences decisions in AEC e-commerce, and emphasizes the need for evaluating each bidder individually. A rating system could support such evaluations.

3.1.3.2.4 Buildpoint

Buildpoint is a former construction e-market place that has restructured to become "the leading Provider Relationship Management (PRM)" [70] to construction companies. It currently provides a subcontractor qualification tool and is in the process of developing an intra company rating tool which was launched in April 2002. Buildpoint's rating tool will be customizable, allowing each contractor to specify the criteria to be rated and their associated weight. As a result, it will be difficult to aggregate information across organizations, since each company can have their own metrics for evaluating subcontractors. Buildpoint's tool calculates overall ratings by aggregating ratings from several raters belonging to the same organization. However, the overall ratings attribute the same weight to all raters.

3.1.3.2.5 Ratingsource

RatingSource is a provider of "unbiased rating information of past and current performance [71]." Its primary product, "Owner Selection," targets public officials and provides evaluations of AEC service providers (mainly general contractors). The evaluation process starts with Ratingsource contacting the service provider, which provides a complete list of its clients. Ratingsource then contacts the clients, who rate the service provider's performance on twenty-five criteria, using a one to ten scale. To compensate for possible erroneous ratings, the overall ratings provided by Ratingsource "are based on a statistically significant sample of a service firm's clients[71]." When calculating the overall ratings Ratingsource assigns the same weight to all raters. Another product of RatingSource is a "Monitoring Database", which general contractors can use to evaluate and track the performance of subcontractors, as well as other AEC service providers. An organization can choose to use the monitoring database internally, but may also share the information with other organizations [71]. The Monitoring Database can also provide the user with different types of reports, such as performance tracking over time and projects. As for Owner Selection, the Ratingsource methodology assigns the same weights to all raters when aggregating ratings in the Monitoring Database.

3.1.3.2.6 Struxicon

In 2000, the AEC electronic market-place start-up Struxicon launched an application to provide background checks of participants in their market place. They offered their members "searches of company licenses, finances, liens, litigation and business practice" [72]. The customer could access and search this information according to customized search criteria. Struxicon used a feature developed by *NEXIS* that allowed "searchable access to over 3.5 billion documents from hundreds of thousands of sources"[73]. However, it did not analyze and integrate the information. A user could obtain the license documents and credit ratings of a contractor, but would have to read these documents for themselves. Similar to many other AEC e-commerce start-ups, Struxicon did not survive the 2001 e-commerce crash.

3.1.4 Summary and discussion of current practice

My investigation of the current practice for evaluating subcontractors in the AEC industry shows that subjective information provided by peer contractors is important when evaluating subcontractors. It also shows that there exist informal mechanisms in place for sharing this information, but also that these are inefficient, since they are either oral or paper based. Another observation is that it is important to consider the source of the information aggregating ratings from peer practitioners.

I have also shown that the Internet provides the opportunity to formalize information about subcontractor performance in AEC. Table 3 shows the online rating applications in current practice that are most relevant to this research project, along with the aggregation methods that they apply. None of the online rating applications, which exist in the AEC industry, considers the source of the ratings when calculating overall ratings. Outside AEC, Open Ratings, and epinions.com do base the weight of a rating on the identity of the rater, but it is not clear how to easily implement these systems in AEC. As shown later in this chapter Open Ratings' solution requires a large amount of data to function, while it is not clear that epinions.com "Network of Trust" approach is consistent with the rationale of AEC practitioners.

Table 3 Existing online rating applications

Name	Type of Application	Rating Aggregation Method
Ebay	Rating system for consumer-to-consumer electronic commerce transactions	Sum of ratings
Bizrate	Ratings of electronic commerce vendors	Unweighted Average
Epinions	Experts evaluating consumer products and services.	Network of Trust methodology
Open Ratings	Rating of suppliers on B2B e-commerce portal	Complex black box formula incorporating Network of Trust as well as Collaborative filtering methodologies
Ralacon and Godkjenningsordningene	Scandinavian contractor prequalification systems	N/A
EASY (NCC)	Internal rating application developed by large General Contractor	Rules
BuildPoint	Internal Rating application for general contractors	None
RatingSource	Rating of general contractors by owners	Unweighted Average
Struxicon	Information about suppliers on e-commerce portal	None

Based on the above discussion, I conclude that this research project has the potential of providing important practical implications to the AEC industry. By implementing and testing the added value of a rating system which takes into account the identity of the rater, this research project will provide insight into how better to design AEC rating systems. This investigation will provide knowledge for the design of rating systems, and covers, for example, personalization, the feasibility of information sharing between organizations, the applicability of including subjecting ratings, as well as the influence of subjective ratings and different rating mechanisms on bid decisions. This insight will benefit contractors, providers of AEC rating solutions, and electronic market place providers. Ultimately, better designed rating systems can contribute to the creation of the trust required for widespread adoption of electronic commerce in AEC.

3.2 Theoretical Point of Departure

The theoretical point of departure for this research project, which investigates how source credibility can support electronic bidding in the AEC industry, covers research from several disciplines. The beginning of the theoretical point of departure will discuss the importance of rating systems from a transaction cost perspective and show why the procurement of AEC services is an appropriate focus for this research project.

I will then continue by discussing earlier work in the four research areas of AEC electronic commerce, AEC bidding, Reputation mechanisms in electronic commerce, and Source Credibility theory, which I identify as directly relevant to this project.

First, I will discuss research in construction engineering and management, which focuses on the emergent area of AEC electronic commerce. Next follows an overview of relevant research in AEC bidding, which is a more mature area of study within construction engineering and management. For the purpose of this dissertation, I have classified research in the area of AEC bidding into two categories: research investigating the rationale of decision makers, and research proposing decision support tools. Having covered construction engineering and management, the next section discusses research which investigates rating systems in electronic commerce. As in the section about AEC bidding, I have classified this research into two categories: research investigating the added value of research rating systems in electronic commerce, and alternative methodologies for constructing weights in rating systems.

The final relevant area of research is source credibility theory. I will first give a general introduction to source credibility theory, before discussing research which investigates the concept's applicability in commercial as well as online settings. The section ends with a discussion of the research that suggests that factors other than source credibility can affect the weight of a rating.

This chapter ends with a summary of this research project's opportunities to make contributions to the stare of research in the four research areas.

3.2.1 Impact of rating systems from a transaction cost perspective

In this section, I will discuss the impact of rating systems in AEC from a transaction cost perspective. I will introduce transaction cost theory before showing that the impact of rating systems will be the greatest for a market characterized by hybrid governance, which is a mix between a free market and hierarchy. As a result, I conclude that it is appropriate to focus this investigation on the subcontracting of services in the AEC industry, which are typically organized using a hybrid governance structure.

3.2.1.1 Transaction cost theory

In his 1937 article "Nature of the Firm" [74], Coase posits that the structure of a firm is set up to minimize the overall transactions costs. Firms should conduct internally only those activities that cannot be procured more cheaply in the market. As a result, a firm will expand precisely to the point where "the costs of organizing an extra transaction within the firm becomes equal to the costs of carrying out the same transaction by means of an exchange on the open market." Building on Coase work, Williamson [75] sums the governance (transaction) and production costs to measure the performance of a governance structure. He also defines three separate governance structures:

- **Market** refers to the free market where each activity is performed by a separate firm. In the construction industry, the procurement of commodities, such as lumber, follows a pure market governance structure.
- **Hierarchy** is the structure when the activities are vertically integrated within one organization. An example of a hierarchy in AEC is a specialized design-build contractor who has internalized the design

function. Other contractors choose to perform critical subcontracts such as structural steel internally.

- **Hybrid** is an intermediate situation between markets and hierarchies. Williamson identifies various forms of long term contracting, reciprocal trading, and franchising as examples of hybrid markets. In the construction industry hybrid is the typical mode in which contractors subcontract services. Eccles [51] has shown that, even though, subcontractors and contractors are legally independent business entities the participants tend to form close long-term relations, a relationship that almost constitutes a "quasi-firm."

The total costs associated with markets and hierarchies include production as well as transaction costs. However, following Williamson's analysis [76], the following analysis will disregard the impact of production costs and focuses on transaction or governance costs. Winter [77] categorizes transaction costs into three types:

- **Frictional costs** of transacting include search costs, ex ante bargaining costs, meetings, contract costs etc.
- **Transactional Hazards** is the focus of Williamson's research and include quality shortfalls, ex-post bargaining over surplus, litigation, hold-up costs, and wasted investments. They are caused by the simultaneous presence of two pair of factors: bounded rationality – uncertainty/complexity and opportunism – small numbers. When the transaction is complex and the outcome uncertain, the bounded rationality of human beings prevents us from making rational decisions. Opportunistic behavior is more likely when there are a small number of actors in the market since competition between a large number of actors generally decreases opportunism. To conclude, transactional hazards are acute when a small number of opportunistic market participants threaten to take advantage of the uncertainty/complexity of the transactions.

- **Opportunity cost of foregone organizational arrangements** is the inability of independent actors to generate the efficiencies we associate with information sharing and collaboration within a single organization.

Williamson concentrates on the second type of transaction and shows how it is a function of asset specificity. Asset specificity occurs when investments required by the two parties involved in a transaction cannot be redeployed to alternatives uses outside the specific transaction. Williamson differentiates between six types of asset specificity 1) site or locational specificity, 2) physical asset specificity (e.g., a specialized crane) 3) human asset specificity or "learning by doing", 4) brand name capital, 5) dedicated assets in the form of discrete investments specific to the relation with a particular transaction partner), and, 6) temporal specificity. In the construction industry, brand name capital (4) and discrete investments (5) are less applicable, while the remaining four types of asset specificity are commonplace. When asset specificity is high one or both parties are "locked into" the transaction. An extreme example of physical asset specificity in the AEC-industry is the customized exterior wall paneling of the museum of Bilbao. The custom-made wall panels would have little value on the open market. An example of locational asset specificity is a job site where there is only one company that can deliver ready-mixed concrete within an acceptable delivery radius [78]. The general contractor would then have the choice of making the concrete on-site or face a possible "hold-up" situation where the concrete manufacturer can take advantage of being a monopoly. Temporal specificity is similar to Thompson's [79] definition of sequential interdependence. Thompson distinguishes three types of interdependences between activities or organizational entities:

- **Pooled interdependence**: Each activity takes place independently but the success of the overall organization relies on that each activity is adequately performed. The failure of one subcontractor may threaten the profitability of the overall project, which may affect all project participants.

- **Sequential interdependence:** Two activities are interdependent in a specified order. On a construction project, form work and concrete reinforcement are examples of sequentially interdependent activities.
- **Reciprocal interdependence:** Two activities are reciprocal if the people involved are mutually and concurrently dependent on one another for information. One example is a fast-track design build project where the contractor analysis the constructability of the drawings. Thomsen et al [80] argues that reciprocal interdependence increases that if the activities' contributions to joint requirements interact negatively. From an organizational design perspective, Thompson [79] puts forward the existence of reciprocal interdependence as major reason for the formation of hierarchies.

Furthermore, Thompson states that the three types of interdependences form a Guttman type (or cumulative) scale. Reciprocal interdependence requires sequential interdependence, which, in turn, requires pooled interdependence.

Increased asset specificity favors hierarchies at the expense of markets. Market governance is suited for goods with low asset specificity while vertical integration is applicable to transactions with high asset specificity.

3.2.1.2 Transaction cost theory in AEC

Gunnarson & Levitt [78] argue that an optimum governance system in the construction industry can be viewed as a function of asset specificity. A typical construction project comprises a large number of transactions of varying asset specificity. The typical governance structure for the construction industry is hybrid. Table 4 gives an overview of major types of procurement in the construction industry and their associated asset specificity. The procurement of commodity products, such as lumber or kitchen appliances, involves little or no asset specificity. The relationship between the buyer and the seller is at an arm's length basis where the buyer, for each transaction, chooses the supplier with the best trade off between product and price. Certain services, such as painting, can also be seen as commodities. However, there is still an element of asset specificity as the buyer is constrained by the availability of service providers in the local market. The governance structure is therefore a mix of hybrid and market. Other services are somewhat more specialized and the transactions give rise to human asset specificity. For instance, key personnel at a general contractor and a controls subcontractor can leverage the lessons learned from working together on one project to the next project. A general contractor often employs the same subcontractor repeatedly, which gives rise to the "quasi-firm" [51] hybrid structure. Subcontracts, who are on the critical path, such as structural steel, also involve temporal asset specificity of sequential interdependence (or reciprocal dependence if the dependence is unidirectional). Schedule interdependence has empirically been shown to be very important source of asset specificity in the ship-building industry [81] which is very similar to AEC. In view of the potential impact of these trades on overall project profitability, some general contractors have opted to internalize them. Another example of an hierarchical organization of highly asset specific transactions is marine subcontractors which purchases their own high capacity floating cranes to avoid potential hold-up [78]. For highly specialized trades, we can therefore categorize the typical governance structure as in between hybrid and market. Finally, in the case of integrated product design and construction the asset specificity becomes even higher. For a specialized design-build general contractor the design is a key determinant of project performance in terms of time, cost, as well as quality. Furthermore, it is difficult to specify and measure the quality of the architect's design until construction is completed. As a result, it is not surprising that many specialized design-build contractors choose to internalize the design.

Table 4: Asset specificity in the construction industry

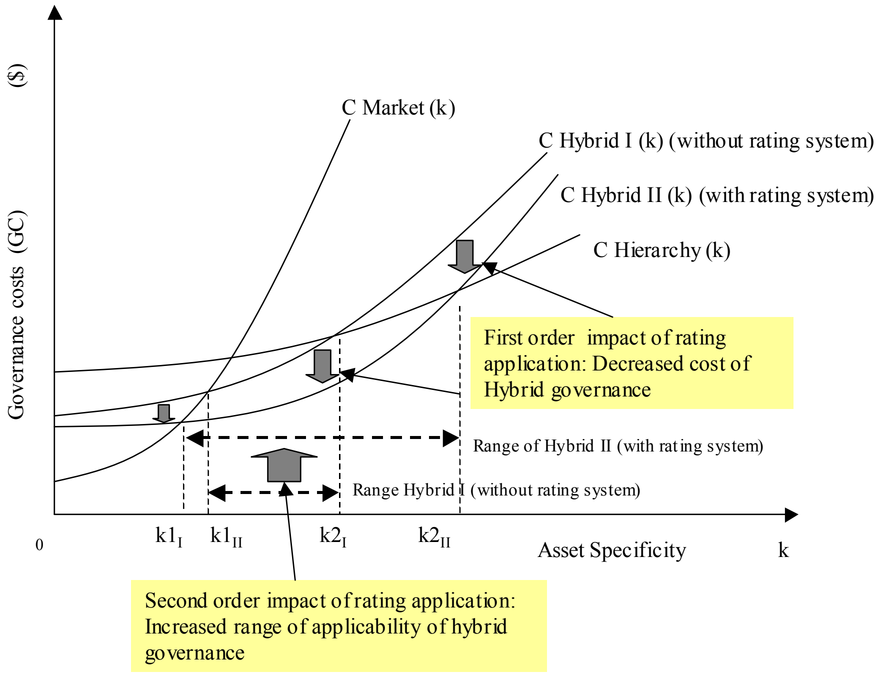
Type of AEC procurement	Commodity Products	Commoditized services	Somewhat specialized services	Specialized services	Integrated product design and construction
Examples in AEC transaction	Procurement of lumber	Sub-contracting of Painting	Sub-contracting of Controls	Sub-contracting of Structural Steel	Specialized Design/Build Contracts
Asset Specificity	Low	Medium/Low	Medium	Medium/-High	High
Sources of significant asset Specificity	None	Locational	Locational Human	Locational Human, Temporal Physical	Locational Human, Temporal
Inter-depen-dences	None	Pooled	Pooled	Pooled Sequential	Pooled Sequential Reciprocal
Typical Governance Structure	Free Market	Hybrid/-Market	Hybrid	Hybrid/-Integrated	Hierarchy (Internal Organization)

3.2.1.3 Predicted impact of a rating system on the governance structure in the construction industry

Williamson [75] identifies the possibility to create an internal feedback mechanism as one advantage that an internal organization enjoys over market governance. He argues that incentives for dishonesty would make it difficult to put such systems into place across organizational borders. In this research project I propose a rating system, which is specifically designed to account for the varying credibility of raters from within and outside the user's organization. I therefore hypothesize that a credibility-weighted rating system can add value also in a hybrid market. The following analysis will show that the introduction of a rating mechanism is of particular interest for the hybrid governance structure.

Figure 4 illustrates the impact of the introduction of a rating application in a construction industry setting. It shows the governance cost associated with the three different types of governance structures as a function of asset specificity (k). Specifically, C Hybrid I (k) and II represent governance costs of the hybrid market structure before and after the introduction of a rating system. The introduction of a rating mechanism could affect two of the three types of transactions costs, which Winter identifies [77]. Firstly, it decreases the frictional costs by reducing the time and effort required to investigate the performance of a potential replier. More importantly, a rating system can affect the transactional hazards by enhancing the reputation effects. In addition, Williamson [76] argues that reputation effects will have the greatest impact in a hybrid market where the transactions involve a medium to high degree of asset specificity. In this setting, reputation effects can "attenuate incentives to behave opportunistically." As a result, ceteris paribus, the impact of improved reputation effect will lead to increased hybrid contracting relative to hierarchies (moving equilibrium from $k2_I$ to $k2_{II}$ in Figure 4 [2]). In AEC, a general contractor may choose to form a partnership with a structural steel subcontractor instead of performing the structural steel in-house. In addition, given that the rating system measures subjective criteria such as collaborativeness and litigiousness, we would expect the resulting reputation effects to be greater in a hybrid relative to a free market. As a result, equilibrium between the Market and Hybrid structure will shift to the left from $k1_I$ to $k1_{II}$ in Figure 4. The enhanced reputation effects associated with a rating system could, for example, make a concrete supplier willing to assume the risk of setting up a just in time production arrangement in cooperation with a major customer. Hold-up would be less attractive to both parties in the transaction given the potential negative impact on their reputations.

[2] A rating tool may shift the position of the cost curves for all three types of market structures. However, since the shift will be greater for the hybrid case, I only show the hybrid curve shifting. This representation is tidier and show the same results as a diagram where all curves are moving to different degrees.



First order impact of rating application: Decreased cost of Hybrid governance

Second order impact of rating application: Increased range of applicability of hybrid governance

Figure 4: Predicted impact of the adoption of rating applications in AEC. The introduction of rating applications will firstly decrease the cost of hybrid governance by strengthening reputation effects. As a result, the range of transactions for which hybrid is the most efficient type of governance increases. (Adapted from Williamson [82])

3.2.1.4 Conclusion

As Figure 4 shows, we can expect two major impacts as the result of the adoption rating system in the construction industry.

Primary Impact: Decreased Cost of Hybrid Governance – The transaction costs associated with hybrid governance will decrease as a rating system reduces frictional transaction costs, as well as strengthens reputation effects which alleviates transactional hazard by reducing opportunism.

Secondary impact: increased hybrid governance – Figure 4 illustrates the range of applicability of hybrid governance before (Range Hybrid I) and after (Range Hybrid II) the introduction of a rating system. As the figure shows, the introduction of a rating system is likely to increase the use of hybrid governance. This increase will primarily be at the expense of hierarchies but can also cause a shift away from market governance.

Going back to Table 4, *this investigation focuses on the subcontracting of AEC services (ranging from commoditized to specialized services).* The investigation will therefore cover transactions with a medium level of asset specificity and for which the governance structure is typically hybrid. I conclude that by focusing on the subcontracting of services, this study will cover the type of transactions where we can expect a rating system to make the most substantial difference from a transaction cost perspective.

3.2.2 Research in AEC electronic commerce

Research focusing on AEC electronic commerce is a relatively new field within construction engineering and management. However, in recent years, researchers have used a number of different perspectives to investigate how electronic commerce can impact and support processes in the AEC industry.

First of all, there is the supply chain perspective. Kim et al propose how agent-based electronic markets can be used for supply chain coordination [83] while Taylor and Björnsson discuss how Internet based pooled procurement can improve the construction supply chain [84]. Both papers argue that Internet technologies make it possible to obtain substantial gains in efficiency through better supply chain coordination.

In terms of Internet strategy, De la Hoz and Ballester Munuz [85] present the prospect of a "business internatization" of the construction industry. More relevant to this project is perhaps Clark et al's [86] study of the strategic implementation of IT in construction companies. They found that construction companies show a strong commitment to both the importance and use of IT to support supplier management. In addition, Koivu [87] found in a Delphi study that, when procuring building services, AEC decision-makers should take into account the value added by the information and benefits to the life cycle of a facility.

The recent proliferation of project extranets in AEC has caught the attention of many researchers. One area of study is the impact of project extranets on document management. Barron and Fischer [88] found that Internet based project control systems could lead to substantial cost savings through facilitation of the processing of documents such as time cards, change orders and monthly billings.

Björk [89] predicts that document management systems will, in the future, be integrated with other ASP services such as bidding. Other researchers have studied the managerial implication of the introduction of project Extranets. Howard and Pedersen [90] monitored the impact of Internet technology on the communication in partnering projects. They showed that in crisis situations project participants often reverted to traditional means of communication [90]. Another case study by Suchocki [91] shows that successful deployment of a project extranet application requires the consideration of the needs of an AEC market participant's entire supply chain. Murray and Lai [92], on the other hand, promote the creation of contract-specific web sites, rather than the use of existing commercial packages. A different approach was taken by Mortensen, who used an ethnographic methodology to study the effects of project management software on project-wide communication patterns [93].

Another relevant topic in the research area of AEC electronic commerce, which has attracted the attention of researchers, is the development of information standards that support AEC electronic transactions (e.g., [94], [95],[96], [97].) Zarli and Richaud [98], for example, investigated the requirements for a standardized open infrastructure, and showed how distributed object systems could support this type of infrastructure. In another study using an AEC electronic commerce value added services perspective rather than a transaction automation perspective, Arnold [99] proposes a framework for automating engineering services over the Internet.

In Europe, the cross-disciplinary PROCURE [100] project demonstrates a methodology for IT deployment through pilots in the LSE (Large Scale Engineering) industry through three parallel pilot projects. The project's focus is on deploying information and communication technology support for collaborative working, product data sharing and knowledge re-use.

The work in AEC electronic commerce that is most relevant to this project is Zolin et al's [1] and Tseng and Ling's [2]. Zolin et al studied trust development in AEC cross-functional global teams. In their original model [1], they identify "gossip and third party information" as important sources of trust. The rating model proposed in this project is intended to formalize and improve the sharing of "gossip and third party information" which would in turn lead to better trust decisions in Zolin et al's trust model. Linking subcontractor evaluation to the Internet, Tseng and Lin present a tool that calculates the utility for each bidding subcontractor based on information stored in XML documents. Ratings weighted by credibility, as proposed in this research project, could complement Tseng and Ling's subcontracting decision support model, which does not differentiate the subcontractors in terms of quality. Instead, their model requires that all subcontractors fulfill the pre-defined quality requirements.

In Table 5 below I summarize important research in the field of AEC electronic commerce in terms of the problems the research addresses, as well as the research methodology applied. The table shows that there is little research focusing on rating systems that supports decision-making in AEC e-bidding or other electronic commerce transaction in AEC.

The table also shows that most research has taken one of two approaches. The first approach is to propose new models or tools that support electronic commerce transactions. The other approach is to use to investigate how existing technologies can add value, for example, through case studies. Few researchers have used experimentation to evaluate the applicability of new technologies that support AEC e-commerce processes.

Table 5 Important research in AEC electronic commerce

Researchers	Area of Study	Research Methodology
Tseng and Ling [2]	Subcontracting decision support based on online information	Modeling
Zolin [1]	Trust in global functional teams	Study of behavior of student teams
Koivu [87])	Added value of information and benefits to the life cycle	Delphi-Study
Leonard et al [100]	Deploying information and communication technology to support AEC processes	Implementation of Pilot projects
Barron and Fischer [88]	Internet Project Control Systems	Case Study
Kim et al [83]	Subcontractor Coordination through Intelligent Agents	Modeling/Charette
Howard and Pedersen [90]	Impact of Internet technologies on communication behavior in AEC projects	Case study
Arnold [99]	Integration of Engineering Services	Modeling
Shreyer and Shwarte [96]	XML standard for building materials	Modeling

3.2.3 AEC bidding

A more mature area of study within construction engineering and management is the study of AEC bidding. This section will first discuss research investigating the rationale for bidding decisions, before presenting research proposing tools that support AEC bidding.

3.2.3.1 Research investigating rational for bidding decisions

In this section, I discuss research that investigates or formalizes the rationale which underlies the bidder's (as well as the evaluator of the bid's) cost or profit estimates. Bidding is a topic that many scholars in construction management have covered. This literature review concentrates on research that focuses on bidding as a decision problem for the individual market participant. However, in this context it is important to mention Tendering theory, or competitive bidding strategy, which is an important body of literature in construction management, but which is not directly relevant for this research project. Tendering theory applies game or decision theory, taking into account not only the bidder's own cost and profit estimates, but also the competitors' strategies, to determine the bid price which will maximize a bidder's profits. (See, for example, [101] for an introduction, or [102] for an overview.) However, the in-depth analysis of competitors' strategies is not the focus of this research project.

More relevant is empirical and descriptive research analyzing decisions of industry practitioners during the bid process. When studying the evaluation of contractors from an owner's perspective, Birrell [3] differentiates between "past experience" and "past performance." Past experience is similar to project experience, which is described in Section 3.1.2.1.1, and identifies the number type and location of projects that the contractor has complicated. Past performance, on the other hand, refers to whether the projects were executed successfully, and therefore involves subjective judgment. Russell et al [103] found that, in the process of prequalifying general contractors, owners place importance not only on objective criteria such as past experience and the number of completed projects, but also on criteria which can only be measured subjectively by peer industry practitioners, such as "Change Order Frequency," "Schedule Performance," and "Willingness to Resolve Conflicts and Problems." Similarly, Holt et al [5, 7] found that the contractor's "actual quality", as well as cost and schedule overruns, influenced the owner's choice of contractors. While it is apparent that there is a subjective element involved in evaluating a contractor's "actual quality," they acknowledge that the same applies to the two latter measures by noting the necessity to "determine what percentage of such overruns are attributable to a firm's failings."

Another area of study in AEC bidding is the general contractor's decisions. From the general contractor's perspective, studies have shown that contract size and type [104], as well as the owner's reputation [9], influenced the bid to the owner. From the perspective of a contractor evaluating subcontractors, Kale and Ardti [105] have shown that contractor performance and profit are both positively correlated to the quality of contractor's relationship with their contractors. They also point to operational and managerial weaknesses as important causes of the high failure rate of small companies (i.e., subcontractors) in the US construction industry.

Gilbane [106] and Shash [107] both evaluate bidding decisions from the point of view of the subcontractor. The two studies [106, 107] showed that subcontractors rely on subjective judgment in order to decide whether to bid or not to bid, as well as how to determine the mark-up that the subcontractor may add to its bids. More specifically, Shash [107] found the general contractor's payment habits to be the single most important factor affecting the amount of mark-up that a subcontractor adds to bids.

Table 6 Research investigating how AEC decision makers evaluate of market in the context of bidding.

Researchers	**Study Characteristics**	**Findings relevant to added value of rating mechanisms in AEC e-commerce**
Birrel [3]. Holt et al [5, 7] Russell et al [103]	Questionnaire to construction owners	Subjective criteria are important for owner's evaluating general contractors
Drew [104]	Statistical analysis of bid for public work	Contract size and type influence general contractors' markup of their bids to the owner
Wanous [9].	Interview and survey targeting general contactors	The owner's reputation influence general contractors' bid to the owner
Kale and Ardti [105]	Survey of General contractors' relationship to when subcontractors	Contractor performance and profit are both positively correlated to the quality of a contractor's relationship with their subcontractors
Gilbane [106] Shash [107]	Survey investigating subcontractor's rationale for bids to general contractors	Subcontractors vary their bids depending on the perceived qualities of the general contractor

Table 6 summarizes the research investigating the rationale underlying decisions in AEC bidding. The studies show that a market participant's perceived quality on subjective criteria, that are measured subjectively by peer industry practitioners, influence decisions in AEC bidding. However, there is no research investigating the impact of the sources that supply the subjective information. The identity of the sources becomes more important as the Internet makes it possible to share information, not only within, but also across organizations. As I have shown in the practical point of departure section, AEC practitioners do rely on subjective information from sources outside their organization to support the current practice of evaluating subcontractors. Therefore, an interesting field of study, especially from an electronic commerce perspective, is how practitioners can take advantage of subjective information from sources of varying reliability to support bid decisions.

3.2.3.2 Tools supporting AEC bidding decisions

Construction management and engineering researchers have proposed an array of tools and models which support bid decisions. The tools and model differ both regarding the methodologies they apply, as well as regarding the perspective they are taking (in terms of who is evaluating whom.)

Based on the work of Friedman [108] and Gates [109] several studies have proposed quantitative mathematical models to support bid decisions. However, due to the difficulty of accurately modeling the input that the calculations require, "these mathematical models proved to be suitable for academia but not for practitioners [9]." As a result researchers have applied methodologies, which partly address this problem, such as Multi-attribute utility theory, Fuzzy set theory, and the Analytical hierarchy process, to bid decisions in construction.

Multi-attribute utility theory (MAUT) calculates the overall utility of an object by computing the weighted average of the object's utility scores on a set of value dimensions [110]. Holt et al [5, 7] have applied MAUT in models that aid the owner in choosing the best contractor. Other applications of MAUT in construction management include the selection of procurement method [111] and the prequalification of contractors [4]. To support contractor's bid decisions Wanous et al proposes a *"parametric solution"* [9], which shares many similarities with MAUT, and computes a bidding index by aggregating the ratings on a set of subjective criteria. Another approach is the analytical hierarchy process (AHP), which applies an eigen-value methodology to calculate the weights that serve to aggregate different criteria [112]. Sik-Wah Fong and Kit-Yung Choi [8] apply a model based on the analytical hierarchy process to help owners evaluate general contractors. Fuzzy set theory, which I will describe in section 3.2.4.2.6 below, can also be used to support AEC bid decisions. Okoroth et al [6] present a model that applies fuzzy sets to analyze subcontractors from a general contractor perspective. In addition, subcontractor risk is a part of Tah and Carr's [113] fuzzy logic based model for construction project risk assessment.

MAUT, AHP and fuzzy sets are not the only methodologies applied in tools that support AEC bidding processes. To help the decision maker select subcontractors for design/build projects, Palaneeswaran and Kumaraswamy [10] present a step-wise evaluation model, which involves the aggregation of multiple criteria. Chinyio et al [11] apply multi-dimensional scaling techniques to map client need to contractor capabilities. Elazouing and Metwally's D-SUB application [114] applies linear programming to minimize a general contractor's total cost when determining what work-items to subcontract or self-perform. Linking subcontractor evaluation to the Internet, Tseng and Lin [2] present a tool that calculates the utility for each bidding subcontractor based on information stored in XML documents. The model is built on the assumption that all of the contractors are pre-qualified.

Table 7 Research in Construction Management proposing tools support bidding

Researchers	**Problem Addressed**	**Evaluation Methodology**	**Strategy for dealing with subjective information**
Tseng and Lin [2]	Subcontracting decision support based on online information	Utility Theory	Subcontractors assumed to be prequalified
Sik-Wah Fong and Kit-Yung Choi [8]	Owners evaluation bids from general contractors	AHP (Analytical Hierarchy Process	Not addressed (only internal ratings are considered)
Palaneeswaran and Kumaraswamy [10]	Owner selecting contractor on Design/Build Projects	Stepwise evaluation model	Not addressed (only internal ratings are considered)
Holt et al [5, 7]	Owner evaluating contractors	MAUT (Multiple Attribute Utility Theory)	Not addressed (only internal ratings are considered)
Chinyio et al [11]	Evaluation of Contractors	Multi Dimensional Scaling Techniques	Not addressed (only internal ratings are considered)
Russel et al [4]	Owner prequalifying contractors	MAUT	Not addressed (only internal ratings are considered)
Wanous et al [9]	Contractor evaluating owner to determine best mark-up on bid	"Parametric solution" (similar to MAUT)	Not applicable (Ratings entered by decision-maker)
Okoroth et al [6]	General Contractor evaluating subcontractors	Fuzzy Logic	Not Addressed (Assumes all experts are credible)

Table 7 summarizes research in construction management, which proposes tools that support bidding in the construction industry. The table shows that little attention has been given to the problem of evaluating subjective information provided by peer industry practitioners using subjective measures. The table shows four strategies, which earlier studies have applied to avoid addressing this problem. The system can assume that all are prequalified (e.g., Tseng and Ling [2]), or that any differences between qualified subcontractors will only marginally impact decisions. Another approach is to limit the tool to internal use within an organization (e.g., Russel [4]). Assuming that all raters within the organization are equally knowledgeable and trustworthy, all ratings should then be equally important. Moreover, Okoroth et al [6] make the same assumption, even when raters from external organizations participate in the system. Finally, Wanous et al [9] let the decision-maker herself input the ratings, and thus avoid any problem when qualifying subjective ratings. As the above analysis shows, there is an opportunity to contribute to the state of research in AEC bidding by investigating the applicability of a rating system that calculates rates depending on the source. More specifically, there is an opportunity to research how source credibility can support rating tools in AEC – bidding.

3.2.4 Reputation Mechanisms in Electronic Commerce

Rating mechanisms, or reputation mechanisms, that support electronic commerce transactions have been an area of study for researchers from several disciplines including computer science, economics, social science, management science, and psychology. To reflect the research questions of this project, I have classified this body of research into two categories. I will first discuss research that has investigated the added value of rating systems, before addressing the issue of operationalization by discussing alternative bases for rating systems.

3.2.4.1 Research analyzing added value of rating mechanisms in electronic commerce

The successful deployment of rating systems in consumer-to-consumer electronic commerce has generated substantial interest in academia. Several researchers have used transaction data from eBay to investigate user behavior in online bidding. The most extensive study to date was Reznick and Zeckhauser's [15]. Reznick and Zeckhauser, who had access to all transactions that took place at eBay for a period of six months, found that the ratings at eBay were "well beyond reasonable expectation" almost always positive, and that buyer and seller ratings were heavily correlated. Another interesting finding was that, contrary to expectations, sellers with high ratings did not enjoy higher prices, although they were more likely to sell their items. However, several other researchers obtained results which contradict this last finding. Analyzing eBay auctions for collector stamps, Dewan and Hsu [13] found that a seller's rating mechanism did positively affect the prices of the stamps sold, but that the effect was marginal. They also found that consumers seemed willing to pay higher prices when using, Michael Roger's, a competing Internet auction site with better trust services, than they would when using eBay. This result indicates that the type of trust mechanism deployed can affect decision-making behavior in electronic commerce transactions. In another study of eBay auctions, House and Wooders [14] showed seller ratings to have a statistically significant impact on prices. Similarly, Lucking-Reiley et al [12] found that a seller's rating has a measurable, although small, effect on the auction prices on eBay. They also found that negative ratings have a much more significant effect than positive ratings do.

Applying game theory to an analysis of the economic efficiency of rating mechanisms like eBay's, Dellarocas [115] found that it is possible to construct an "optimal judgment" rule. If all users are sophisticated enough, and have the information available to make optimal judgments, it is, in theory, possible to arrive at a steady state where buyers accurately predict the sellers' true quality. However, he concludes that, in practice, this is probably not what takes place at auctions like eBay.

Using data from epinions.com, Chen and Pal Singh [18] compared the performance of their proposed solution, which applies the concept of reputation hierarchies (see section 3.2.4.2.5), and the epinion rating mechanism. They show that reputation hierarchies enable the classification of raters into two groups, "Good" and "Bad," where the "Good" raters' ratings are more consistent. They also found that people tend to trust active raters, even if their rating quality is not high. Using data from Yahoo's auctions, they found that buyers and sellers' ratings were significantly correlated, and that an overwhelming majority of the ratings were positive. This finding is consistent with Reznick and Zeckhauser's [15]. One reasonable explanation for it is "that people give high ratings to others in the hope of getting high ratings in return" [18].

There are also a few studies which do not rely on online transaction data for their analysis. List and Lucking-Reiley [50] performed field experiments where they auctioned sports cards, and found that buyers seemed to behave more rationally for high-priced than for lower-priced cards. From a Human Computer Interface (HCI) perspective, Swearingen and Sinha [116] performed an experiment to evaluate four different commercial rating systems. They found that an effective rating system should have a logic that is "at least somewhat transparent."

Presenting the preliminary results of case studies, Ratnasingham and Kulmar [16] found that trust in business-to-business electronic commerce can be seen from two perspectives: "trading partner trust as between human actors in e-commerce" and "technology assurances (as in trust and security based mechanisms in e-commerce)." They also note that "while ample research exists in the case of the latter perspective, only limited research exists in the role of trust between human actors." This research project, serves to bridge this gap by proposing and evaluating a model that formalizes trust between human actors to support rating mechanisms.

Table 8 Research investigating the added value of rating systems in electronic commerce

Researchers	**Study Characteristics**	**Findings relevant to added value of rating mechanism**
Reznick and Zeckhauser's [15]	Statistical analysis of eBay transaction data	A seller's rating do not affect price but affects likelihood of the item being sold
Dewan and Hsu [13], Lucking-Reiley et al [12], House and Wooders [14]	Statistical analysis of eBay transaction data	A seller's ratings have a small but statistically significant impact on price.
Dellarocas [115]	Game theoretical analysis of eBay transaction data	Theoretically possible to arrive at a steady state where buyers use 'optimal judgment' to predict sellers' true quality
List and Lucking-Reiley [50]	Field experiment auctioning sports cards	Users behave more rational when buying high-priced items
Chen and Pal Singh [18]	Statistical Analysis of Data from Epinions, and three auction sites	Users trust active raters more. Possible to classify rater into "Good" and "Bad"
Swearingen and Sinha [116]	Experiment testing HCI of four commercial rating systems	Transparent logic is important
Ratnasingham and Kumar [16]	Case study of B2B electronic applications	Trust between human actor is relevant in B2B e-commerce

Table 8 summarizes the research that has investigated the added value of rating mechanisms in electronic commerce. As shown, most research has focused on consumer-to-consumer electronic commerce while little research has studied the added value of rating mechanisms in the context of B2B e-commerce. There is also little research comparing the effects of different rating mechanisms on decision-maker behavior during electronic transactions.

3.2.4.2 Alternative Bases for e-commerce rating systems

The following sections will present seven different bases for e-commerce rating systems. The different bases have all been proposed by researchers but do not include source credibility, which I will discuss in the following section. In particular, I will discuss how each of these bases can support an online rating system of AEC subcontractors. This section ends with a summary comparison of the pros and cons of the seven different methodologies.

3.2.4.2.1 Network of Trust

Milgram [117] showed that two randomly chosen persons within the United States could be linked by six or fewer first-name references [48]. This observation inspired the so-called law of “Six Degrees of Separation” as it appears in Guare’s book of the same title [118]. The advantage of this approach is that it “enables a user to leverage the knowledge of several users […] to find services of the desired type and of high quality without excessive communication [119].” Zacharia and Moukas [65] operationalize a “Network of Trust” in *Histos,* a reputation mechanism designed for “ highly connected online communities.” *Histos* is a personalized rating mechanism where all ratings are individual and depend on what actor is considered and from whose perspective. User A’s rating will be different from B’s and C’s perspective, respectively. The aggregate rating of user A from user C’s perspective depends on the direct path of ratings between them. The underlying assumption is that trust is transitive or, in other words, that people tend to trust the friend of a friend more than someone unknown [120]. For example, if B rates A highly, and C rates B highly, then the rating of A from C’s perspective will be high as well. The strength of a “Network of Trust” is that it can build upon existing relationships. This could be important for an AEC e-commerce community moving from an online to an offline presence. . One advantage of a rating mechanism based on a network of trust is that it would be easy to “cold start”, or to start from scratch. To get up and running, the system simply requires one set of users to enter whom they trust. Furthermore, Ono et al [121] have applied a network of trust based models to construct an e-commerce trust application using Java Agents. Their model is general enough to handle any B2B setting, but in their study they apply it to AEC subcontractors bidding for jobs. While the Histos, model involves complex calculations Ono et al’s model is much simpler, which creates a more transparent system from a user perspective. The two applications both use a single dimension to model trust, which reflects a party’s trustworthiness as a business partner (Will he cheat in business?) and credibility as an evaluator (Can I trust what he is saying?). Furthermore, this trust or credibility dimension is measured using an arbitrary scale. As mentioned in section 3.1.3.1.3, the attitude towards the idea of a network of trust varies considerably among AEC decision makers. For industry practitioners, the concept of trusting “a friend of a friend” seemed intuitive, while others did not consider it to be relevant if they and an unknown rater turned out to have a common friend.

3.2.4.2.2 Collaborative Filtering

A collaborative filtering system measures the extent to which users agree in terms of taste. The underlying assumption is that if two users agree about the quality of item A they will also agree about item B. The underlying reasoning can be exemplified as follows: If both Sam and Alice like the movie *Titanic*, and Sam also likes the movie *Lord of The Rings*, then it is likely that Alice also will like *Lord of The Rings*. Pioneering work applying collaborative filtering in Internet based recommender systems was done by Reznick et al in the Group Lens project [17] and Shardanand and Maes [122] who proposed collaborative ratings mechanisms to automate the “Word of Mouth.” Collaborative filtering can be expanded to incorporate, for example, content-based filtering [123], item-based filtering [124], “Network of Trust” dimensions [65], and case-based reasoning [125]. Collaborative filtering has been found to be most applicable when a large set of users rates items that are difficult to quantify and describe (e.g., movie ratings, books) [48]. My investigations have shown that subcontractor performance can efficiently be measured using a handful of predetermined criteria (See 3.1.2). The major problem with collaborative filtering is that it requires a substantial amount of data to obtain useful results [124]. This may not be a problem for e-commerce merchants such as Amazon but could pose substantial difficulties in AEC e-commerce, especially during a start-up phase. Furthermore, it is not certain that the weights calculated by a collaborative filtering mechanism are consistent with user expectations. It may very well be that a user's rating behavior is more similar to that of an unknown rater than to that of a trusted friend. In this case, the unknown rater would enjoy a higher weight than the trusted friend, even though this is probably not consistent with the user's personal beliefs.

3.2.4.2.3 Rule-based mechanisms

Abdul-Rahman et al [22] proposes an algorithm which uses rules to determine and update rater weights. They recognize that trust is context specific (e.g., Bob may trust Alice in the context of repairing bikes but not regarding brain surgery) and they also include mechanisms to gradually tune the user's assessment of whom to trust based on the outcomes of previous interactions. In their paper, they propose rules to guide the agents choices of who to trust. The problem with this approach is that the rules tend to be ad-hoc. For example, if Chuck believes that Jim, an estimator at CalGC's, has rated of PaintA is inaccurately, how much would his trust in Jim decrease?

3.2.4.2.4 Statistical Analysis of Past Ratings

Statistical analysis of past ratings can serve to evaluate the raters in an AEC e-commerce market. Avery et al. [126] proposed this solution to support an Internet market for consumer evaluations. They suggest that a statistical analysis will provide a means of monitoring the quality of subjective information such as ratings. The idea is that someone whose ratings deviate substantially from those of the rest of the market may not be honest or may not possess the expertise to rate properly. However, as Avery et al point out, if one only penalizes raters who are too far from the mean, this may discourage raters from posting idiosyncratic opinions and result in a universal “average” rating. In their article, they suggest penalizing raters who deviate too much from the average, as well as those whose ratings are too close to the average. Another way of measuring this property is to calculate two separate variance measures. The first is the within subject variance, or the variance calculated over the set of ratings a rater has contributed. A high within subject variance indicates that the rater does differentiate between the different items he or she is rating. The second measure is the ‘between-subject variance’, or the mean squared distance between a rater’s ratings and the average rating, which is calculated for the entire set of raters. A high between subject variance indicates that a rater’s ratings are either very idiosyncratic (and hence less likely to be of interest to other raters), or that the rater is dishonest. Dellarocas [19] takes a more sophisticated approach, using clustering too differentiate between dishonest and honest raters. He argues that, given certain assumptions about steady-state and user strategies, cluster filtering can serve to significantly reduce the impact of fraudulent behavior. I argue that the main problem with using a statistical analysis to analyze rating behavior is that, no matter how sophisticated the filter is, a fraudulent user who knows how the system works will probably always be able to come up with a strategy for being dishonest without being detected. Similar to collaborative filtering, there is also a risk that weights calculated through statistical analysis will be inconsistent with user expectations.

3.2.4.2.5 Reputation Hierarchies

Chen and Pal Singh [18] propose a reputation hierarchy as a means to explicitly calculate rater reputation. The strength of this model is that it adjusts for a rater’s expertise, which may vary depending on the domain that is being rated. Their system also calculates the confidence level of each rating. The model calculates weights through a propagation of endorsement across a hierarchy of raters and groups. The endorsement level is a function of the discrepancy between raters and groups, who have rated the same item. They also claim that their methodology can help limit the impact of dishonest raters, since a rater needs a good reputation to have an impact on the overall aggregated ratings. As for collaborative filtering, a major problem with reputation hierarchies is that they need a substantial amount of data to calculate the weights. Reputation hierarchies also run the risk of presenting ratings that are inconsistent with user expectations.

3.2.4.2.6 Fuzzy Sets

Fuzzy logic is a superset of conventional (Boolean) logic, which has been extended to handle the concept of "partial truth." Truth values lie between "completely true" and "completely false" [127]. As its name suggests, the logic's underlying modes of reasoning are approximate rather than exact. Pioneering work in fuzzy logic was done by Zadeh [128], who built on the fact that most modes of human reasoning, and especially common sense reasoning, are approximate in nature.

Zimmerman and Zysno [129] and others [20],[130] have applied fuzzy set theory to formalize the credit rating processes which banks conduct. AEC subcontractor evaluation is a similar process, and it is therefore likely that fuzzy sets can also be applicable in this context. However, I identify three major problems of applying fuzzy logic to an AEC e-commerce rating system

Choice of Operators - One problem with the fuzzy set approach is that its success depends on what operators are used when aggregating the information, a choice which is often arbitrary.

Converting Input to fuzzy numbers – Another difficulty of applying fuzzy sets is capturing input data and converting it to a fuzzy membership function. For example, does the fact that a contractor has completed 70 projects mean that they should be assigned a membership function for the set "very good" of 0.7?

Integration of Credibility – There is no natural way of integrating the impact of source credibility into a fuzzy set rating system. One solution would be to make the ratings fuzzier the less credible the source is. But again, this solution would necessitate arbitrary judgments. The designer would have to choose what operator to use to "fuzzify" the ratings depending on the credibility of the source.

To summarize, it is far from evident how these three problems can be overcome in a fuzzy set rating system in AEC. Furthermore, by moving from real to fuzzy numbers, the system reaches another level of complexity. This will possibly lead to more accurate results, but also results in a system where output becomes more difficult to interpret and validate.

3.2.4.2.7 Subjective Probabilities

In my investigations, I have explored the possibility of constructing a rating system based on subjective probabilities. This is in line with Gambetta's [131] definition [22] of trust in terms of subjective probabilities. A rating model based on subjective probabilities would incorporate the four following principles:

1) A rating should express both the expected overall value of the rated criterion, and the uncertainty which is associated with this overall value.
2) The system calculates the overall rating as a weighted average.
3) The *credibility* of the ratings is inferred from the actors' assessments of each other's credibility, and contributes to the variance of a rating.
4) The systems determine the *weight* of each rating so as to minimize total variance of the overall rating. This way uncertain ratings, which contain a lot of noise, will be discounted, but they will still contribute to the overall value of the ratings.

One way to operationalize the model would be to follow the Howard's [132] 5-step interview process (Motivating, Structuring, Conditioning, Encoding, and Verifying.) One problem with this solution is that it is far from certain that trust is always equal to subjective probabilities [133]. Moreover, the model is based on the assumptions that all the parameters conform to normal distributions. Finally, any implementation would involve the assessment of a very large number of parameters, which is likely to be difficult, given the time pressured environment of AEC decision-makers.

3.2.4.2.8 The applicability of alternative methodologies as a basis for a rating mechanism in AEC e-bidding

Table 9 summarizes the strengths and weaknesses of the seven different rating methodologies as a basis for a rating system in AEC e-bidding. Studying the different methodologies, we observe that there are three major approaches to calculating weights in rating systems. Each of the three approaches has major problems that have to be overcome for a methodology to be applicable to AEC e-bidding. In Network of Trust and Subjective Probabilities, the user directly enters whom he or she trusts. The problem is that the measuring of the input that the models require is either difficult (e.g., network of trust) or is likely to be very time consuming (as with subjective probabilities). Collaborative filtering, reputation hierarchies and statistical analysis all calculate the weights from transaction/rating data. The problem with this approach is that the calculation of accurate weights requires a lot of rating data, which would be difficult to obtain in an AEC market place. Finally, rule based mechanisms and fuzzy sets rely on ad-hoc operators to aggregate information. I therefore argue that there is an opportunity to research alternative bases for methodologies to aggregate ratings in B2B electronic commerce transactions. The following section presents source credibility theory, which constitutes a promising basis for a rating system, overcoming many of the problems of existing methodologies.

Table 9 The strengths and weaknesses of different methodologies as a basis for rating mechanisms in AEC e-bidding

Principle	**Strength**	**Weakness**
Network of Trust [21, 121]	Easy to cold-start	Difficult to measure trust of second degree
Collaborative Filtering [17, 122]	Do not require assumptions about user preferences	Requires substantial amount of data for calibration More applicable for consumer e-commerce
Rule based mechanism [22]	Differentiates between trust in a rater and trust in a buyer/seller	Would rely on ad-hoc rules
Statistical Analysis [19, 126]	Can filter out dishonest raters	Requires substantial amount of data for calibration Always possible for sophisticated user to trick the system
Reputation Hierarchies [18]	Explicitly calculates rater reputation	Requires substantial amount of data for calibration Difficult to cold start.
Fuzzy Sets [20, 134]	Explicitly deals with uncertain information	Would rely on ad-hoc complex operators
Subjective Probabilities	Straight forward statistical calculations	Difficult to measure all required probabilities

3.2.5 Source Credibility Theory

The primary purpose of this section is to introduce source credibility theory and to explain why it is a promising basis for a rating system in AEC e-bidding. I first introduce source credibility theory before discussing research investigating its applicability in commercial and online settings. This discussion also points out the opportunity of further research to explore the applicability of source credibility in online commercial settings. The section continues by presenting factors other than source credibility which are likely to affect the weights of ratings.

3.2.5.1 Introduction to source credibility theory

The importance of *ethos*, or, more commonly, *source credibility* as a means of persuasion has been noted since classical times. A common point of departure of most modern researchers [135] is Aristotle's *On Rhetoric* [136], where ethos is defined as persuasion through character, or a means of persuasion that makes the speaker worthy of credence [136]. Fogg and Tseng [41] point out that trust and credibility often are confused in academic literature. They provide the following useful distinction between the two concepts, which, even though similar, are fundamentally different:

trust <- "dependability"

credibility <- believability or "trust in information"

Most contemporary writers agree that source credibility is multi-dimensional and can be defined as "the attitude toward a source of communication held at a given time by a receiver" [23]. Hovland, Janis and Kelly [24] performed pioneering work, identifying perceived trustworthiness and expertise as the main dimensions of a source's credibility. The higher the trustworthiness and expertise a source is judged to have, the higher the importance given to information coming from that source.

One illustrative example, used in Hovland's initial studies in the `50s, relates to the then still open question, "Can a practical atomic-powered submarine be built at the present time?" To answer this question, Hovland et al [24] present Robert J. Oppenheimer, the inventor of the nuclear bomb, as a credible source. They also argue that most Americans would see the Soviet newspaper Pravda as a less credible source. The American public perceived Oppenheimer to be trustworthier, as well as superior in terms of expertise, relative to this subject. Several researchers [33, 35, 37] have developed scales to measure the construct of source credibility. (Section 4.2.2.1, in the research methodology chapter, provides a thorough discussion of the different scales.) Using scientifically validated scales mitigates the problem of measuring the input that confronts, for example, the Network of Trust rating methodology.

3.2.6 Source Credibility in Commercial Settings

Early work applied and validated source credibility theory in the context of public opinion (e.g., [24, 37, 137]) and interpersonal communication [37]. Later studies have shown that source credibility applies also to commercial settings, where the receiver of information is evaluating possible transactions. Examples are found in areas such as advertising [27], evaluations of used cars [25], job offer acceptance [138], buy or lease decisions [139] job performance evaluations [42], admitting students to business schools [26], and price-perceived risk relationships [28]. These studies all support the applicability of source credibility in commercial settings, but also present more elaborate models, studying factors such as message framing ([42], [28], [139]) and [138]), multiplicative rather than additive models in conjecture with the source' bias [25], time pressure [26], and order of presentation [32], as well as the impact of factors other than trustworthiness and expertise [27]. Although, it is possible that these more elaborate models are applicable to an AEC rating application, the proposed operationalization of the concept, which the research methodology chapter (4) presents, comprises only the basic elements of the construct. This limitation serves to focus the research on the primary research objective, which is to investigate whether source credibility can be used to construct the weights in an AEC rating application. Once we have determined the validity of this hypothesis, we can investigate more complex models to calculate the weight. One example of arguments in favor of a more complex, expanded model relates to the interaction between source and message, even though the message, in this case, is minimal (a rated number). For example, a user may regard a negative rating from a source whom the user expects to be positively biased [138], such as a supplier to a rated subcontractor, as very important.

Table 10 Research investigating source credibility in commercial settings

Investigator	Situation	Aspect Studied
Birnham [25]	Buying Used cars	Bias
Fischer et al [138]	Job offer acceptance	Message Framing
Albrigth and Levy [42]	Buy or lease decisions	Rating Discrepancy
Harmon [139]	Job performance evaluations	Message Framing
Higgins [26]	Admitting students to business schools	Time
Grewal et al [28]	Price-perceived risk relationships	Message Framing

Table 10 summarizes research which investigates source credibility in commercial settings. As Table 10 shows, the task of evaluating subcontractor performance shares similar characteristics with the situations studied in earlier research. I argue that the evaluation of AEC subcontractors is a commercial setting where: 1) decision-makers can gain from sharing information; 2) there are incentives for deceit; and 3) the sources of information (the raters) are likely to have varying expertise. As a result, investigating whether source credibility theory is applicable in this setting constitutes an interesting research opportunity.

3.2.6.1 Source Credibility online

Several researchers have developed models which incorporate elements of source credibility theory to evaluate online information. In a literature review of the credibility of medical websites, Constiantides and Swenson [140] discuss both the theoretical point of departure as well as practical applications. They show that existing efforts to rate and qualify medical websites have been undertaken by professional organizations (e.g., American Medical Association, [141]), and academic institutions (such as Emory University [142], as well as commercial entities (Healthwise [143], for example). Out of a large body of work discussing the evaluation of online medical information, Siberg et al's work [30] is one of the most frequently cited [140]. They identify four core standards (*authorship, attribution, disclosure*, and *currency*) that are available to a user who is judging the quality of a medical web site. Of these standards, authorship is the dimension which is closest to the general definition of source credibility. Furthermore, experimentation has shown

[144] that both knowledge of content and source expertise affect perceptions of online health information.

Another body of research treats website credibility in a more general setting. Comparing different types of media, Flanagan and Metzger [145] found evidence that people consider information obtained online equally credible to that obtained from television, radio, and magazines, but less credible than newspapers. Critchfield [31] found a relationship between the user interface design of a web site and its perceived credibility. In a large quantitative study, Fogg et al [29] found seven factors which determine the perception of web credibility. Five of these dimensions (e.g., "real-world feel") do not apply to raters of subcontractors, but the remaining two are "expertise" and "trustworthiness." The results are therefore consistent with Hovland's original model, as well as with Fogg and Tseng's [41] argument that the two key dimensions of computer credibility are perceived trustworthiness and perceived expertise.

Table 11 Research investigating applicability of source credibility in online settings

Setting	Researchers
Websites in General	Critchfield [31], Fogg and Tseng's [41], Flanagan and Metzger [145], Fogg et al [29]
Online health information	Siberg [30], Eastin [144], Constiantides and Swenson [140]

Table 11 shows that, even though several studies have demonstrated the applicability of source credibility to evaluate online information sources, there is a lack of research investigating the applicability of source credibility in the context of B2B electronic commerce. It is not clear that industry specialists making online purchasing decisions will judge information in the same manner as consumers evaluating online health advice do. More specifically, there is no research into the incorporation of source credibility into rating systems which support electronic commerce transactions.

3.2.6.2 Factors other than credibility affecting weight of ratings

Several researchers have shown that source credibility is not the only factor that comes into play when users weight information from difference sources.

In their work on reputation mechanisms, Zacharia et al [65] point out that *time* is an important factor in determining the weight of information. Similarly, research in communication shows that a source's perceived credibility varies over time [26, 33, 34]. In interviews, I have found that AEC estimators often regard ratings that are more than two years old as substantially less credible than recent ones, even though the discount factors seemed to be highly individualized.

Stone and Stone [146] reported that people perceived information from two sources as more credible than information from a single source. As a result, a potentially useful indicator to the user of a rating system is the *total rater credibility*. The user will want to know whether only one rater with low credibility has rated a subcontractor, or if the overall ratings are based on ratings from several credible raters. One way to calculate such a measure is to total the credibility of all raters who have rated a given subcontractor.

Stone and Stone [39] also found *feedback consistency* to be an important indicator in the context of performance feedback from multiple sources. Similarly, Albright and Levy found [42] that *rating discrepancy* affected recipients' reactions to ratings. One way to model consistency is to use an agreement index that calculates the extent to which the ratings of a subcontractor are consistent.

Grewal, Gotlieb and Marmorstein [28] and others (e.g., [138], [139]) identified the importance of *message framing* in the context of risk assessment. Initial investigations indicate that some AEC decision makers tend to attribute greater importance to negative ratings than to positive ones. One way to account for this effect is to model each user's risk preference.

Newhagen and Nass [40] point out the importance of differentiating between an *organization* and an individual when evaluating the credibility of a message source. My interviews with industry practitioners show that this is certainly true for the construction industry. An estimator may regard a GC as being a not very reputable company, but still trust those of the GC's employees whom he knows on a personal basis. Research undertaken by Mackie et al [147] shows that receivers tend to find "in-group" people more credible than people belonging to "out-groups." This finding underlines the importance of taking into account the rater's organizational affiliation. If the user does not know anything about a rater, she is likely to find the rater more credible if the two belong to the same organization. Furthermore, evaluating the source credibility of a rater's organization enables a rating system to incorporate ratings from a wide variety of sources. The system can use a rater's organization as a proxy if the user does not know the rater. This approach will decrease the time and effort required to capture user references, the input of the system.

Table 12 Factors, other than source credibility, which impact weight of information

Impact on Perceived Credibility	**Researchers**
Importance of Organization	Newhagen and Nash [40], Mackie et al [147]
Number of Sources	Stone and Stone [146]
Feedback Consistency	Stone and Stone [39], Albrigth and Levy [42]
Time	Applebaum et al [34], Andersen [33], Higgins [26]
Message Framing	Grewal et al [28], Fischer et al [138], Harmon [139]

Table 12 summarizes factors, other than source credibility, which can affect the weight of ratings. This research project investigates the importance of three of these factors: organization, number of sources (total credibility), and feedback consistency. The reason for not investigating the impact of time and message framing is threefold. First of all, it is already clear from earlier research that time and message framing are important when aggregating information. Secondly, time and message framing are factors, independent of the credibility of the sources, making their impact of lesser interest in this study. Thirdly, incorporation of these two factors would increase the complexity of the model, possibly confuse the users testing the application, and thus make analysis of user behavior more difficult.

3.2.7 Using Source Credibility Theory to aggregate information from different sources

The goal of this research project is to investigate how to apply source credibility to calculate the overall rating of a subcontractor. Source credibility theory provides a useful basis for aggregating information from multiple sources. The simplest, and most common, information aggregation model in source credibility theory literature is the weighted average (see, for example [148], [32]). Birnhaum et al [25], on the other hand, propose a multiplicative model, even though they do not disprove the appropriateness of a weighted average model. In this research model, I have chosen to use the weighted average model, given that it is the least complex and most extensively tested of the two models. I do not believe that the possible gains of applying a multiplicative model will outweigh the increased complexity that would result. In addition, a multiplicative model would be additive on a log scale.

The statistical method of Bayesian inference provides another rationale for the use of a weighted average model. Bayesian inference weights observations by their "precision." Assuming that the variables are normally distributed, an observation's precision equals the inverse of its variance [149]. If we assume that source credibility models the precision of an observation/rating, a weighted average model will then be consistent with Bayesian inference. Using the weighted average model, a rating system based on source credibility would rely on operators to aggregate information which are to a much lesser extent ad hoc than the operators in rating systems based on Rules Based Mechanisms and Fuzzy Set theory.

3.3 Source Credibility a promising bases for AEC rating mechanisms

In section 3.2.4.2, I presented three major criticisms of the applicability of existing rating mechanisms to AEC e-bidding: 1) reliance on input parameters that were difficult to measure, 2) reliance on ad hoc operators, or 3) the requirement of large datasets of rating/transaction data for calibration. We will now show how a rating system based on source credibility has the potential to mitigate all the of these problems

First, source credibility theory provides tested frameworks (e.g.,[25]) for aggregating ratings from different sources. These frameworks decrease the dependence on ad-hoc operators. Second, there are validated scales for measuring a source's (rater's) credibility [35]; these can serve as the key input parameter in a rating system based on source credibility. Finally, the weights in a rating based on source credibility theory depend on user preferences and not on rater behavior, which decreases the amount of data required to calibrate the rating application. The opportunity to measure the credibility of the rater's organization as well as the credibility of the person further decreases the amount of user input needed. Table 13 summarizes the opportunities for solving the three major problems of existing rating mechanisms, using a rating system based on source credibility.

Table 13 Summary of how a rating system based on source credibility theory mitigates major problems of alternative rating mechanisms

Alternative Methodologies	**Key Problem of deploying Alternative Methodology in AEC e-commerce**	**Opportunity for solution using source credibility based reputation mechanism**
Network of Trust, Subjective probabilities	Difficult to measure input parameters	Scientifically validated scales
Rule based mechanisms, Fuzzy sets	Rely on Ad hoc operators for aggregating ratings	Validated aggregation functions
Collaborative Filtering Reputation Hierarchies, statistical analysis	Need large amounts of clean data for calibration	Relying on user preferences rather than rater behavior decreases the amount of data needed for calibration. Measuring credibility of the organization further decreases amount of user input needed

As shown, there exists an interesting research opportunity to investigate how source credibility theory can support a reputation mechanism in AEC electronic commerce.

3.3.1 Summary of Research Opportunities

Based on the preceding discussion, this section summarizes the opportunities for this research project to contribute to the state of knowledge in four different fields: *AEC electronic commerce, AEC bidding, Rating mechanisms for electronic commerce,* and *Applicability of Source Credibility theory*. The summary below shows that opportunities for contributions exist in all of the four fields of research.

3.3.1.1 Opportunities for contributing to research in *AEC electronic commerce*

In AEC electronic commerce, there is little research investigating the:

- Applicability of Rating systems to support AEC electronic commerce transactions
- Added value of source credibility theory as a basis for rating system in AEC e-bidding.
- Applicability of experimentation to investigate the added value of technologies that support electronic commerce in AEC.

3.3.1.2 Opportunities for contributing to research in *AEC bidding*

In AEC bidding there is little research investigating the:

- Added value of source credibility theory as a basis for a rating system in AEC bidding.
- Impact of the credibility of the sources, which supply subjective information that supports bidding decisions.

Several researchers have proposed tools supporting the AEC bidding decisions but there exists no *AEC bidding tool which assigns weights depending on the source of the ratings.* More specifically, no research has investigated the applicability of a rating tool which calculates rater weights through a methodology that:

- Formalizes source credibility theory
- Incorporates subjective ratings from sources of varying credibility
- Incorporates ratings from raters who belong to different organizations.

3.3.1.3 Opportunities for contributing to research in *Rating mechanisms in electronic commerce*

There is little research in the field of Rating Mechanisms in electronic commerce that:

- Compares the added value of different rating mechanisms in B2B electronic commerce
- Investigates the applicability of source credibility as a basis for a rating mechanism

3.3.1.4 Opportunities for contributing to research investigating the *Applicability of Source credibility theory*

A substantial body of research investigates the applicability of source credibility in commercial as well as online settings. However, there is little research investigating source credibility in an online, *as well as* a commercial setting (i.e., electronic commerce). In particular, there is little research investigating the applicability of source credibility in online situations characterized by 1) substantial benefits from online information sharing, as well as 2) opportunities for deceit.

4 Research Methodology

4.1 Introduction

This chapter discusses the research methodology I have applied to investigate the fundamental research question:

How can source credibility theory support rating systems in the procurement of AEC services?

The two principal research methods of this research project are modeling and experimentation. Upon researching current practice and the theoretical point of departure, I operationalized source credibility into computer model named *TrustBuilder*. I developed two different versions (*TrustBuilder I*, and *TrustBuilder II*), *of* TrustBuilder which served two purposes: 1) investigate the feasibility of operationalizing source credibility, and 2) support experiments (*Experiment I* and *II*) investigating the added value of a source credibility based rating system in AEC. Table 14 summarizes the resulting key research activities in chronological order.

Table 14 Description of key research activities in chronological order

Research Activities in Chronological Order	**Description of Activity**
Design TrustBuilder I	Designed basic subcontractor rating model incorporating source credibility to calculate rater weights
Evaluate TrustBuilder I in Experiment I	Evaluated applicability of using source credibility to support an AEC rating tool in experiment with non-experts
Design TrustBuilder II	Incorporated lessons learnt in Experiment I when designing refined source credibility based rating model
Evaluate TrustBuilder II in Experiment II	Repeated Experiment I in a setting where experts use refined model to evaluate subcontractors based on actual ratings

The modeling section (4.2) of this chapter covers the basic characteristics of the TrustBuilder tool, before describing each version in detail. As shown in Table 14, TrustBuilder II was a refinement of the original TrustBuilder I model and incorporated the lessons learnt from testing TrustBuilder I in Experiment 1.

This chapter continues with a discussion of the two experiments, which studied the two models in the context of AEC e-bidding. Both Experiment I and Experiment II were within-subject designs, with rating tool as the differentiating factor and user as the primary unit of analysis. The two experiments mainly compared the performance (in terms of ability to predict rater weights and added value) of a credibility weighted rating tool (TrustBuilder I or II) to that of a standard, unweighted tool. In Experiment I, non-experts used the basic TrustBuilder I model, which was populated with hypothetical ratings, to evaluate a set of subcontractors. In Experiment II, a set of experts applied the more advanced TrustBuilder II to evaluate a set of subcontractors, the actual performance of which had been rated. The chapter ends with a discussion of the fundamental research hypotheses that the experiments explored.

4.2 Modeling

4.2.1 Introduction

Modeling was an integral part of the research methodology. I designed and implemented two different versions of a rating model which both were based on source credibility. The operationalization served two purposes. First of all, it enabled me to investigate the first part of the research question: *How is it possible to operationalize source credibility to support the calculation of weights that are based on rater identity in an AEC rating tool?* Secondly, a rating tool operationalizing source credibility theory was a prerequisite for conducting two user experiments. These experiments investigated the second part of the research question, which referred to the added value of such a tool. In this section I discuss the two versions of the TrustBuilder tool, which operationalizes source credibility theory. I will first discuss the critical features of the TrustBuilder methodology, which are common to the two versions TrustBuilder I and II. Next follow detailed descriptions of the design of TrustBuilder I, and II. The section ends with a summary comparison of the two tools.

4.2.2 Overview of the TrustBuilder credibility-weighted rating tool

TrustBuilder I and II shared the same basic characteristics. I developed both versions of the tool in Microsoft Excel, applying Visual Basic to design the user interface and code the algorithms for data input as well as calibration. Both versions of the tool provided user directions and recorded user behavior during the experiments. Furthermore, both versions are based on the same three-step process. This process helps the user transform a set of ratings from different raters into information, which supports the evaluation of subcontractors. The three steps are 1) Credibility input, 2) Calculation of rater weights, and 3) Display of ratings and rater information. In each step, TrustBuilder I and II share common design choices. First of all, the tools need a scale to measure credibility. Secondly, the calculation of rater weights requires a calibration procedure. Finally, the tools calculate rater agreement and total rater credibility before displaying these measures.

4.2.2.1 Credibility Input: Choice of scale to measure credibility

The use of source credibility as a basis for a rating system requires a measurement scale. A significant number of studies have strived to develop and refine Hovland et al's [24] original model, which identified perceived Trustworthiness and Expertise as the two dimensions of a source credibility. One of the more cited studies is Berlo et al's [37] which identified three factors Safety, Qualification, and Dynamism, of which the first two are basically equivalent to Hovland's two factors. Guenther [150] did in an overview of factor-studies find considerable inconsistency, which he mainly attributed to the variation of methodologies used and sources studied. Singletary [151], for instance, found as many as 41 factors.

However, Hovland's original Trustworthiness and Expertise reappear in most studies. For example, both Berlo et al, and Vandenbergh [27] models include these two basic factors, even though they argue that additional factors come into play as well. These additional factors are, however, mostly specific to the context of the study and they are therefore not very relevant for an AEC rating system. Berlo et al's Dynamism Factor, for example, relates to how a person presents the information and comprises items such as "frank-reserved", "energetic-tired." As a result, "Dynamism" does not apply to the setting of a rating system where the sources submit their ratings over the Internet using a terminal. The McCroskey [35] Likert scale[3], which has emerged as the predominant measure of source credibility [152], is also very similar to Hovland's original concept. McCroskey's "Authorativeness" and "Character" correspond to Hovland's "Trustworthiness" and "Expertise." Nonetheless, it is also important to mention that in later studies McCroskey [36] has identified Goodwill as a third dimension, which determines source credibility. In addition, Constantinides and Swenson [140] argue in favor of adding goodwill as a third factor in the context of evaluating the credibility of medical web sites. However, a closer examination of the scale measuring the goodwill factor reveals that it is less applicable to electronic bidding. The scale contains items, which either belong to McCroskey's [35] original Character factor (e.g., Selfish) or require that the user knows the source personally (e.g., "Cares about me", "Has my interest at heart"). Asking a project manager to assess whether an estimator at a peer contractor "Has his interest at heart" also creates methodological risks. The participant may become irritated or take the task less seriously. In addition, the design of most rating applications prevent raters from varying the ratings depending on who is the reader/user (e.g., provide Bob with truthful ratings while sending Alice dishonest ratings.) As a result, it becomes less necessary to evaluate the direct relationship between the user and the rater, which is basically what the Goodwill factor does. When evaluating the elements of computer credibility, Fogg [41] also proposes measuring source credibility in terms of the two standard dimensions: Trustworthiness and Expertise. Based on the discussion presented above, and given that the original McCroskey scale [35] has been cited over 100 times; and "seems to have emerged as the predominant method of scaling"[152]; I have chosen to use this

[3] The McCroskey scale is a 12 item semantic differential 7-point Likkert scale. The 12 items have been shown to factor into the two dimensions Authorativeness (Expertise): "Reliable", "Uninformed", "Unqualified ", "Intelligent", "Valuable", "Inexpert"; and Character (Trustworthiness): "Honest‟, "Unfriendly", "Pleasant", "Selfish ‟, "Awful", "Virtuous", "Sinful‟

scale to operationalize the credibility of message sources in TrustBuilder I and II. In order to avoid confusion, I will, in accordance with standard terminology, refer to the two dimensions of the McCroskey scale as Trustworthiness and Technical Expertise, rather than McCroskey's Character and Authoritativeness.

4.2.2.2 Taking into account to what extent the user knows the rater

It is important for an AEC rating system to integrate information from raters who the user knows, as well as from anonymous raters. Moreover, Newhagen and Nass [40] point out the importance of differentiating between an organization and an individual when evaluating the credibility of a message source. My interviews indicate that this difference also applies to people in the construction industry. An estimator may regard a general contractor as being a not very reputable company, but still trust those of the GC's employees whom he knows on a personal basis. Conversely, if the user does not know anything about the rater as an individual, she is likely to account for the organization that the rater works for when assessing the rater's credibility.

In order to enable a rating tool to calculate the credibility for all raters in a rating system, and to take into account the importance of the organization, I therefore define three different situations, which reflect the extent to which the user knows the rater. The user evaluates rater credibility for each of the three cases:

- **Known Rater**: If the user knows the rater, the McCroskey scale can be directly applied to measure credibility.
- **Unknown Rater, Known Organization**: A user is likely to find an unknown rater more credible if the rater works for a well-reputed contractor.
- **Unknown Rater, Unknown Organization**: According to My interviews, AEC practioners differ in the way in which they judge information from "a typical" or "unknown" practitioner in the local industry. Some would never trust the information given by someone they do not know, while others like to "think the best of people." To provide a baseline measure for rater credibility, the user evaluates an unknown rater who works for an unknown organization

on the McCroskey Scale. The two tools can then compare the credibility of known raters and organizations to this baseline measure.

4.2.2.3 Methodology to convert credibility scores to weights

To convert the credibility measures to weights, both TrustBuilder I and II apply a methodology of pair-wise comparisons. This methodology is based on the approach of the Analytic Hierarchy Process (AHP). AHP normally uses an eigen-vector method to calculate the weights of the different criteria, allowing for inconsistencies among the pair-wise comparisons [153]. However, a least-square method generates similar results [154] and facilitates the incorporation of a two level model. Estimating the weights of ratings requires the added complexity of a two level model, and both TrustBuilder I and II therefore apply least-square estimation. Both too also use an exponential target function to ensure that the weights are positive.

We also have to decide at what level the weights in the target function should be calculated. There are three obvious alternatives: 1) the individual user level, 2) company level, and 3) the entire user base. Calculating the coefficients at the individual user level enables the tools to account for each user's unique preferences. However, most users will find it tedious to make more than twenty pair-wise comparisons. Moreover, twenty data points are not always sufficient to obtain stable estimates of the four coefficients, especially if some users are inconsistent when they perform the pair-wise comparisons. Another alternative is to estimate the coefficients at the company level. This approach requires that at least ten users are available from each company and that user preferences within each company are homogenous. The third alternative is to perform the estimation across the entire user base. If all users' preferences were sufficiently homogenous then this approach would warrant stable results. To facilitate experimentation, both TrustBuilder I and II perform the calibration at the individual user level. The tools collect twenty-one data points from each user and estimate the individualized coefficients. The main advantage of this approach was that it enabled the users to participate in the experiment independently of one another. If, on the other hand, the tools were to estimate the coefficients for the entire user group, I would have to split the experiments into two separate parts. In the first part of the experiments, all users would make the pair-wise comparisons. I would then let the tool estimate the weights, before going back to each user a second time, where they use the tool to evaluate subcontractors. This type of approach risks prolonging the research process, making it harder to find industry practitioners who are willing and able to participate[4].

4.2.2.4 Calculation of Rater Agreement

Rater agreement, or the extent to which the different raters agree upon the performance of a given subcontractor, can be valuable information for the industry practitioner who is evaluating bidding subcontractors. In order to calculate rater agreement the two tools apply an adapted version of a raw agreement index [155]. In contrast to standard raw agreement indexes, the TrustBuilder tools incorporate the notion of rater credibility. Alternative methods, such as Latent Class Models [156], Factor Analysis [157], polychoric correlation[158], and calculation of kappa coefficients[159], may in some cases provide more accurate results, but they are also more complex. Since, ceteris paribus, "a simpler statistical method is preferable to a more complicated one" [160], I propose using the relatively simple method presented below.

The idea of accounting for rater credibility is best illustrated through a simple example. Let us assume that that four raters have rated a subcontractor's (PaintA) performance on a scale which comprises the three values High", "Medium", and "Low." The user, who wants to find determine the rater agreement, considers two of the raters (A and B) to be very credible, having estimated their credibility as 9 on a scale from 1 to 10. The user perceives the other two raters (C and D) to be less credible, and has rated their credibility as 5 and 4 respectively. As shown in Table 15, the only two raters who agree on PaintA's performance are A and C, who both rated it as "Medium." From the user's perspective it is natural that the agreement

[4] Nonetheless, when analyzing the results of Experiment II, I pooled the data from the 15 users to calculate the weight coefficients (β_{kr}, $\beta_{so,}$ β_x and β_{tw}) for the entire user base (see Section 5.3.4.1). I could then also analyze the performance of the different rating models using more stable estimate, which are based on a large pool of data.

regarding PaintA’s performance would be higher if the rater who agrees with A about PaintA’s performance were the credible rater B, rather than the less credible C. In order to account for rater credibility, I propose using a raw agreement index where each rater is given as many “votes” as his/her perceived credibility. We can then easily calculate then overall rating agreement by totaling the number of agreeing “votes”, before dividing this sum by the total number of “votes” that could possibly agree. Equation 1 calculates the agreement index for subcontractor j’s performance from user i’s perspective (AI_{ij})

$$AI_{ij} = \frac{\sum_{l}\sum_{k \neq l} A_{jkl} C_{ik}}{\sum_{l}\sum_{k \neq l} C_{ik}} \qquad (1)$$

where:
A_{jkl} equals 1 if raters k and l agree on j’s performance, and 0 otherwise.
C_{ik} equals rater k’s credibility from user i’s perspective.

Table 15 This example illustrates the calculation of Rater Agreement in TrustBuilder I and II. Agreement is calculated using a simple agreement index while adjusting for rater credibility

Rater X	**Rater X’s Credibility**	**Rater X’ Rating of Paint A’s performance**	**De facto Agreement between Rater X and other raters**	**Total Possible Agreement if Rater X agreed with all other raters**
A	9	Medium	5 (Agrees with C)	18
B	9	High	0	18
C	5	Medium	9 (Agrees with A)	22
D	4	Low	0	23
Total Agreement			14	67
Agreement Index = De facto Agreement/Total Possible Agreement = 14/67 = 0.21				

Knowing that the agreement index is 0.21, as in the example, is of course of little value unless there is a benchmark measure to with this result can be compared. One standard way of obtaining such measures is to run a bootstrap analysis, which calculates the mean and variance of the agreement if the ratings were completely random. The user can then complement bootstrap analysis by evaluating a of a set of test cases, where she states whether she perceives the agreement to be “High” or “Low”. However, since the calculation of the agreement measure was not a focus of the investigation, both tools apply a less rigorous conversion method. The user can see a symbolic measure (“High”,”Medium”, “Low”) of the agreement between the raters. For the purpose of the experiments, I assigned the break off values, which differentiates between “High”, ”Medium” and “Low”, in the following way. The tools calculate the agreement for all subcontractors, before ranking them by rater agreement. Finally, the tool divided the subcontractors into three groups of equal size with “High”, “Low”, or “Medium” rater agreement.

4.2.2.5 Calculation of Total Rater Credibility

The user will also be interested in knowing the total credibility of the raters who have evaluated a subcontractor. Are the overall subcontractor ratings based on one unknown rater or twenty trustworthy experts? TrustBuilder I and II calculate the total rater credibility by totalling the credibility of all the raters who have rated a given subcontractor (Equation 2.)

$$TotC_{ik} = \sum_{j} C_{ij} \qquad (2)$$

$$Rjk \neq 0$$

where
TotC$_{ik}$ = Total Credibility from user i’s perspective of all raters of subcontractor k.
C$_{ij}$ = User i’s estimate of rater j’s credibility

To convert the numeric total credibility measures to symbolic measures the two tools apply the same approach as for Rater Agreement. The tools calculate the total credibility for all subcontractors, before ranking them by rater total credibility. Finally, the tool divided the subcontractors into three groups of equal size with “High”, “Low”, or “Medium” total credibility.

4.2.3 TrustBuilder I: A prototype rating tool operationalizing source credibility theory

4.2.3.1 Introduction

TrustBuilder I is a basic version of the TrustBuilder rating tool which operationalizes source credibility theory to calculate subcontractor ratings. The purpose of implementing TrustBuilder I was to support Experiment I, which tested the applicability of source credibility as a basis for a rating system which supports AEC e-bidding. To calculate the aggregate ratings of subcontractors TrustBuilder I follows the three step process of 1) Credibility input, 2) Calculation of rater weights, and 3) Display Ratings and rater information.

4.2.3.2 Step 1: Credibility Input

The user evaluates three different types of raters on the twelve-item McCroskey credibility scale. To compensate for individual behavior and preferences, TrustBuilder I normalizes the scores for the different scale-items to z-scores[5].

Unknown Rater: The user evaluates an unknown person who works for an unknown organization. The tool adds user's z-scores for each item to obtain the credibility of an unknown rater C_u.

$$C_u = \sum_{\substack{Rater\_Unknown \\ Organizati\,on\_Unknown}} z\_scores \tag{3}$$

C_u corresponds to the base line credibility that the user attributes to anybody working in the AEC industry.

Unknown Rater, Known Organization (Organizational Credibility): TrustBuilder I calculates the credibility of an unknown rater working for a known organization in the same way as it does the calculation for a completely unknown

[5] z-scores measures a scale reading's distance from the mean in terms of standard deviations. In this case the mean and standard deviation were calculated for each user and scale item.

rater. The credibility of the organization (C_o) then equals the sum of z-score minus the credibility of the unknown rater (C_u).

$$C_o = \sum_{\substack{Rater\_Unknown \\ Organization\_Known}} z\_scores - C_u \qquad (4)$$

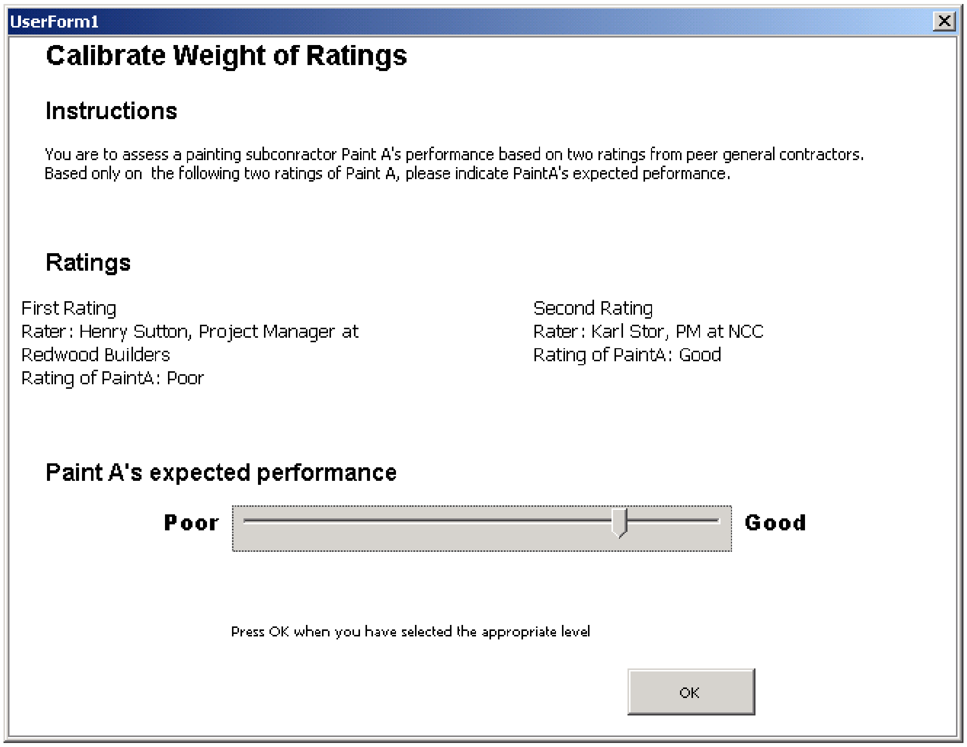
C_o reflects the organizational credibility, or the net credibility attributed to the organization for which the rater works.

Direct Knowledge (Personal Credibility): In this case the user knows the rater and therefore also the organization. The personal credibility of the rater (C_p) equals the sum of the z-scores, minus the credibility of an unknown rater (C_u), minus the credibility attributed to an unknown rater working for the same organization ($C_{o.}$).

$$C_p = \sum_{\substack{Rater\_Known \\ Organization\_Known}} z\_scores - C_o - C_u \qquad (5)$$

4.2.3.3 Step 2: Calculation of rater weights

The next step of TrustBuilder I is converting the three measures of credibility (C_p, C_o and C_u) into a single measure, which represents rater credibility or weight. To carry out this conversion, it is necessary to estimate a constant (α) and two coefficients (β_p and β_o). β_p and β_o represent the relative importance that the user attributes to the rater's personal and organizational credibility. One user may judge that the primary factor that determines a rater's overall credibility is the rater rater's personal credibility, which would imply a large β_p. Another user may regard the rater's organizational affiliation as very important (large β_o), while a third user may regard all raters as equally credible (large constant (α) relative to coefficients). As section 4.2.2.3 described, the two tools use a methodology of pair-wise comparisons to estimate these weights. TrustBuilder I uses a logistic regression function to estimate the coefficients since it provides a better model for rater weight than a normal linear regression. In TrustBuilder I, the user performs the pair-wise comparisons in a user interface where a painting subcontractor (“PaintA”) has been rated by two raters (see Figure 5).



Calibrate Weight of Ratings

Instructions

You are to assess a painting subconractor Paint A's performance based on two ratings from peer general contractors.
Based only on the following two ratings of Paint A, please indicate PaintA's expected peformance.

Ratings

First Rating
Rater: Henry Sutton, Project Manager at Redwood Builders
Rating of PaintA: Poor

Second Rating
Rater: Karl Stor, PM at NCC
Rating of PaintA: Good

Paint A's expected performance

Press OK when you have selected the appropriate level

Figure 5: User interface to calibrate weight of ratings through logistic regression. The user indicates a subcontractor’s expected performance based on two divergent ratings.

The first rater (Rater 1) rated PaintA’s performance as “Good”, while the second rater (Rater 2) rated it as “Poor”. Based on these two ratings, the Participants submit their evaluations of PaintA’s performance by dragging a continuous slide-bar in between the values “Poor” and “Good”. This value (w_{12}) corresponds to the weight that the user attributes to Rater 1’s ratings vis-à-vis Rater 2’s. By modeling each rater’s credibility as an exponential function, we obtain the following model for $\hat{w}_{12}$:

$$\hat{w}_{1,2} = \frac{C_1}{C_1 + C_2} \tag{6}$$

where:

$$C_j = \frac{1}{1 - \exp(\alpha + \beta_p C_{pj} + \beta_o C_{oj}}$$

$$j = 1,2$$

We can then estimate α, β_p and β_o by minimizing:

$$\sum_{k,l} (\hat{w}_{kl} - w_{kl})^2 \tag{7}$$

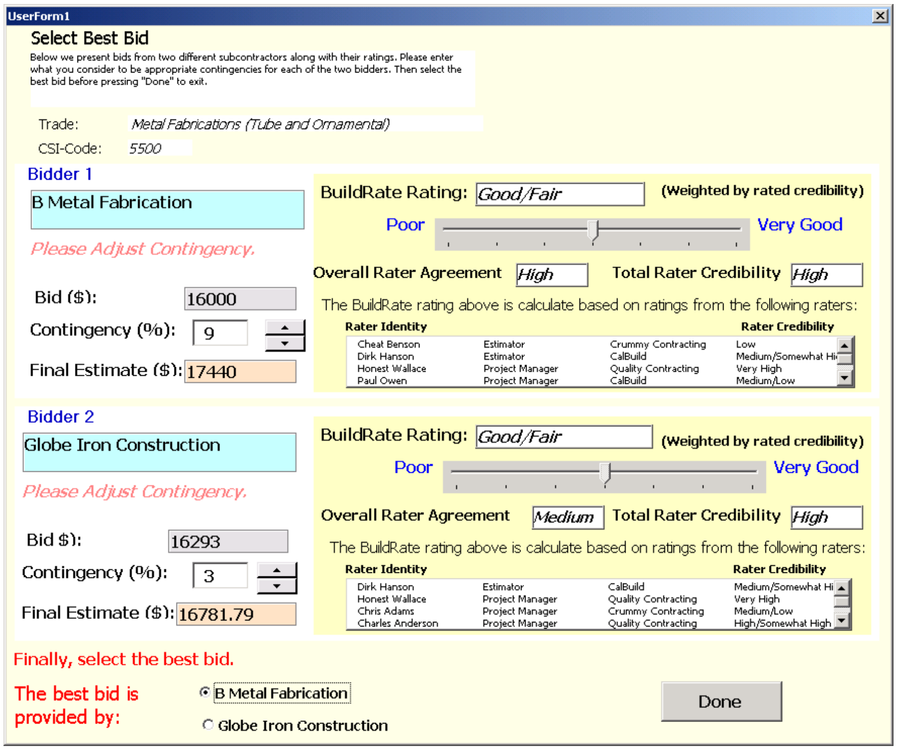
As mentioned above, the resulting estimation function corresponds to a logistic regression. The overall rating R_{mi} of a subcontractor m from the user i's perspective will then equal the ratings provided by each rater j multiplied by the rater's estimated credibility. As a result, we obtain the following straightforward formula:

$$\begin{gathered} R_i = \sum_j R_{ij} * C_j / \sum_j \; C_j \\ Rij \neq 0 \end{gathered} \tag{8}$$

After having calculated a numeric value for the credibility of each rater, the system asks the user to evaluate the raters with the highest and lowest numeric credibility values on a ten-point Likert-scale. TrustBuilder I uses these evaluations to translate the numeric credibility values to symbolic values.

4.2.3.4 Step 3: Display Ratings and rater information

In the prototype user interface the user can see the calculated values for two different subcontractors. An example of the user interface is shown in Figure 6 below. The user can see the overall rating (weighted by credibility) both on a continuous scale and as a symbolic value. She can also see the rater agreement and total rater credibility for each subcontractor, along with the individual credibility of each rater on a symbolic scale. To support Experiment I, the TrustBuilder I prototype also allows the user to input contingency for each bid and select the best bidder.



UserForm1

Select Best Bid

Below we present bids from two different subcontractors along with their ratings. Please enter what you consider to be appropriate contingencies for each of the two bidders. Then select the best bid before pressing "Done" to exit.

Trade: *Metal Fabrications (Tube and Ornamental)*

CSI-Code: *5500*

Bidder 1

B Metal Fabrication

Please Adjust Contingency.

Bid ($): 16000

Contingency (%): 9

Final Estimate ($): 17440

BuildRate Rating: *Good/Fair* (Weighted by rated credibility)

Poor — Very Good

Overall Rater Agreement *High* Total Rater Credibility *High*

The BuildRate rating above is calculate based on ratings from the following raters:

Rater Identity			Rater Credibility
Cheat Benson	Estimator	Crummy Contracting	Low
Dirk Hanson	Estimator	CalBuild	Medium/Somewhat Hi
Honest Wallace	Project Manager	Quality Contracting	Very High
Paul Owen	Project Manager	CalBuild	Medium/Low

Bidder 2

Globe Iron Construction

Please Adjust Contingency.

Bid $): 16293

Contingency (%): 3

Final Estimate ($): 16781.79

BuildRate Rating: *Good/Fair* (Weighted by rated credibility)

Poor — Very Good

Overall Rater Agreement *Medium* Total Rater Credibility *High*

The BuildRate rating above is calculate based on ratings from the following raters:

Rater Identity			Rater Credibility
Dirk Hanson	Estimator	CalBuild	Medium/Somewhat Hi
Honest Wallace	Project Manager	Quality Contracting	Very High
Chris Adams	Project Manager	Crummy Contracting	Medium/Low
Charles Anderson	Project Manager	Quality Contracting	High/Somewhat High

Finally, select the best bid.

The best bid is provided by:

● B Metal Fabrication

○ Globe Iron Construction

Figure 6: User interface showing bids from and ratings of two subcontractors. The user is asked to input contingency for each bid and select the best bid.

4.2.4 TrustBuilder II: A more refined rating tool operationalizing source credibility theory

4.2.4.1 Introduction

Implementing Trustbuilder II fulfilled two purposes. Firstly, Trustbuilder II supports the presentation of more information and hence provides a more realistic environment compared to TrustBuilder I. The expert industry practitioners who tested TrustBuilder II in Experiment II required a more realistic context when evaluating subcontractors. Secondly, based on lessons learnt from implementing and testing Trustbuilder I in Experiment I, TrustBuilder II employs a refined model for estimating credibility. The refinement lies in the factors used to measure rater credibility and in the function which converts credibility measures to weights. Similar to TrustBuilder I, follows a 3-step process to calculate ratings.

4.2.4.2 Step 1: Credibility Input

The set of factors, which measure credibility, is not identical in the two models. TrustBuilder I models rater credibility as a function of three factors (Credibility of Unknown Rater, Credibility of known organization, and Credibility of known Rater). In the TrustBuilder I model, credibility is based on the user's assessment of the credibility of an unknown rater. The model then adjusts this baseline credibility in case the user knows organization or the rater. Experiment I showed that several participants did not use the McCroskey scale in the same manner when they evaluated an unknown rater, as they did when they knew the rater very well. Another conclusion from Experiment I was that the model should differentiate between perceived trustworthiness and expertise. (See factor analysis of Experiment I in Section 5.2.4.1 for more details.) User discussions and pre-testing also demonstrated the importance of accounting for whether the rater and the user work for the same organization. As a result, the TrustBuilder II model comprises four variables: whether the user knows the rater, whether the two of them work for the same organization, rater trustworthiness, and rater expertise. In case the user does not know the rater personally, TrustBuilder II sets the rater's Expertise and Trustworthiness to be those of a typical rater working for the same organization as the rater (if the organization is known), or to be those of a typical unknown rater.

TrustBuilder II uses four different factors to model $C_{ij,}$ or user i's estimate of rater j's credibility:

- **Know Rater (KR_{ij}):** Does user i know rater j? This is a binary measure entered by the user.
- **Same Organization (SO_{ij})**: Do user i and rater j work for the same organization? The model calculates this binary measure based on the two's organizational affiliation.
- **Rater Expertise (X_{ij})**: What is the expertise of rater j according to user i's opinion? Table 16 shows the calculation of X_{ij}.
- **Rater Trustworthiness (TW_{ij}):** What is the trustworthiness of rater j according to user i's opinion? Table 16 shows the calculation of TW_{ij}.

While the "Know Rater" and "Same Organization" variables are easy to model using binary measures, TrustBuilder II applies the McCroskey [35] scale to model Rater Expertise and Trustworthiness. Table 16 shows the scale and its operationalization in TrustBuilder II.

Table 16: The McCroskey scale and its operationalization in the Trustbuilder II rating tool.

Factor	**Scale items**	**Operationalization:**
Authorativeness (Expertise)	Reliable-Unreliable Uninformed – Informed Unqualified – Qualified Intelligent – Unintelligent Valuable – Worthless Expert – Inexpert	$X_{ij} = \sum_{k=1}^{k=6} x_{ijk}$
Character (Trustworthiness)	Honest – Dishonest Unfriendly - Friendly Pleasant - Unpleasant Selfish - Unselfish Awful - Nice Virtuous -Sinful	$TW_{ij} = \sum_{k=1}^{k=6} tw_{ijk}$

Evaluating trustworthiness and expertise is straightforward when the user knows the rater. The user evaluates the rater using the McCroskey scale and TrustBuilder II calculates rater expertise and trustworthiness as described above. TrustBuilder II also calculates rater expertise and trustworthiness in case the user does not know the rater. As in TrustBuilder I, the system therefore has the user evaluate two types of "typical" but unknown raters:

1. "***Typical Project Manager working for Contractor X***": The user rates the expertise and trustworthiness of typical project managers working for each contractor that 1) the user knows, and 2) has supplied ratings to the system. This way the system will have a value to assign for raters who are unknown to the user, but who works for a contractor the user is familiar with.
2. "***Typical Project Manager working for a typical California contractor***": The purpose of having the users evaluate this type of rater is to enable the system to assign expertise and trustworthiness values to raters in case both the organization and the individual are unknown to the user.

The system calculates the overall scores for all raters on the four factors KR, SO, X, and TW, before converting them into z-scores[6]. The normalization ensures that, for each user, all factors will have a mean of zero and a standard deviation of one. As a result, TrustBuilder II can calibrate the credibility model across users, as Section 5.3.4.1 of the results chapter will exemplify.

My objective has been to base the conversion function, which estimates overall rater credibility, on simple and adopted models. The exponential model used in Trustbuilder I, is similar to that of a logistic regression. Trustbuilder II, employs an even simpler exponential function to model rater credibility. The advantage of an exponential function compared to the function used in TrustBuilder I, is that it better models variance[161]. As a result, the proposed function is consistent with Bayesian inference, which weights observations by their precision. If the variables are normally, an observation's precision equals the inverse of their variance [149]. In addition, we can use least square regression to evaluate the exponential function, which also fulfills the constraint of being positive. Equation 9 formalizes the estimate of rater j's credibility from user's i's perspective:

$$C_{ij} = \exp(-1 + \beta_{KR} KR_{ij} + \beta_{SO} SO_{ij} + \beta_X X_{ij} + \beta tw TW_{ij}) \qquad (9)$$

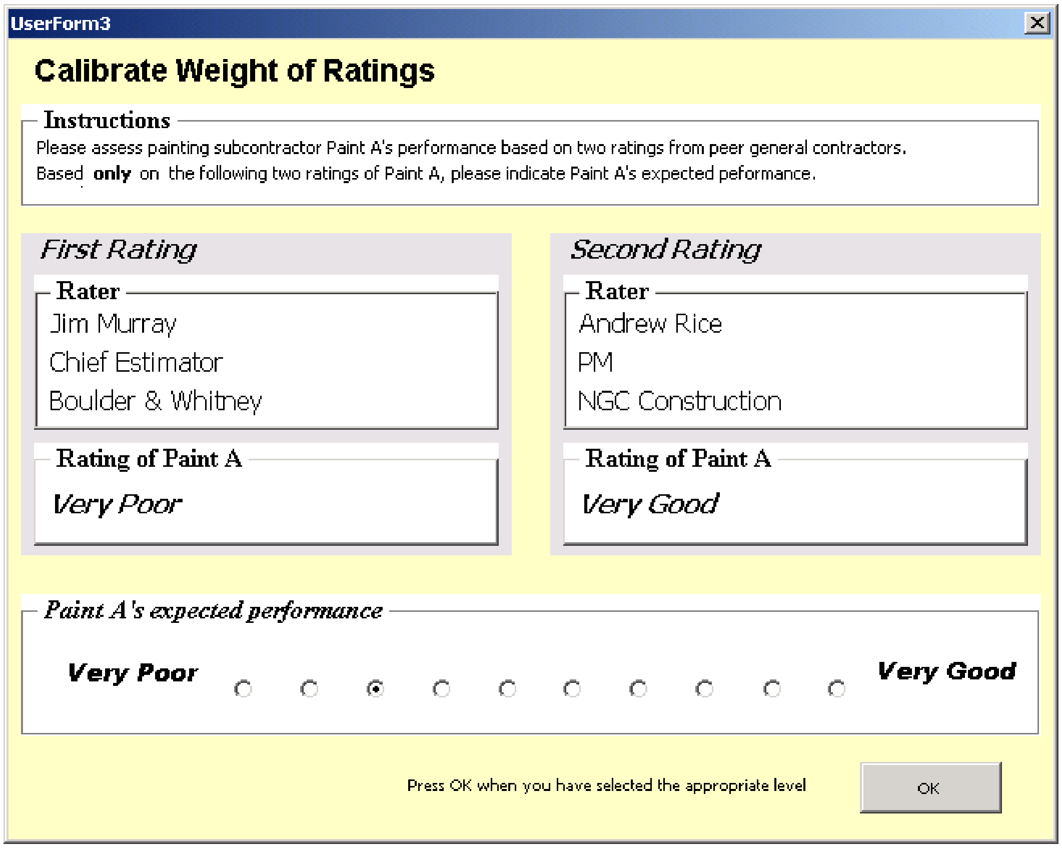
where: KR, SO, X and TW are user i's z-scores for rater j on each of the four factors and β_{KR}, $\beta_{SO,}$ β_X and β_{TW} are coefficients which reflect the importance of each factor.

4.2.4.3 Step 2: Calculation of rater weights

To estimate the coefficients of Equation 9, TrustBuilder II uses a methodology of pair-wise comparisons. The only difference from the calibration procedure of TrustBuilder I is that the user evaluates rater weight on a ten-point Likert scale instead of continuous slide bar. This change occurred since several participants in Experiment I expressed concern over the difficulty of being consistent

[6] z-scores measures a scale reading's distance from the mean in terms of standard deviations. In this case the mean and standard deviation were calculated for each user and scale item.

when using the slide bar. As in TrustBuilder 1, the tool shows a user interface where a painting subcontractor ("PaintA") has been rated by two of the seven raters (see Figure 7).



UserForm3

Calibrate Weight of Ratings

Instructions

Please assess painting subcontractor Paint A's performance based on two ratings from peer general contractors.
Based **only** on the following two ratings of Paint A, please indicate Paint A's expected peformance.

First Rating

Rater
Jim Murray
Chief Estimator
Boulder & Whitney

Rating of Paint A
Very Poor

Second Rating

Rater
Andrew Rice
PM
NGC Construction

Rating of Paint A
Very Good

Paint A's expected performance

Very Poor ○ ○ ● ○ ○ ○ ○ ○ ○ ○ **Very Good**

Press OK when you have selected the appropriate level

Figure 7: User interface to calibrate weight of ratings through pair-wise comparisons.

Rater 1 rated PaintA's performance as "Good" and Rater 2 rated it as "Poor". Participants submit their evaluations by clicking a 10 point Likert scale between the values "Very Poor" and "Very Good". The value (w_{12}) corresponds to the weight that the user attributes to Rater 1's ratings vis-à-vis Rater 2's. By modeling the credibility of each rater as an exponential function, we obtain the following model for $\hat{w}_{12}$:

$$\hat{w}_{1,2} = \frac{C_1}{C_1 + C_2} \tag{10}$$

where:

$$C_{ij} = \exp(\alpha + \beta_{KR} KR_{ij} + \beta_{SO} SO_{ij} + \beta_X X_{ij} + \beta tw TW_{ij})$$

TrustBuilder can then estimate β_{KR}, $\beta_{SO,}$ β_X and β_{TW} by minimizing the sum of squares of the errors associated with all pairs (k,l) of raters included in the pairwise comparisons.

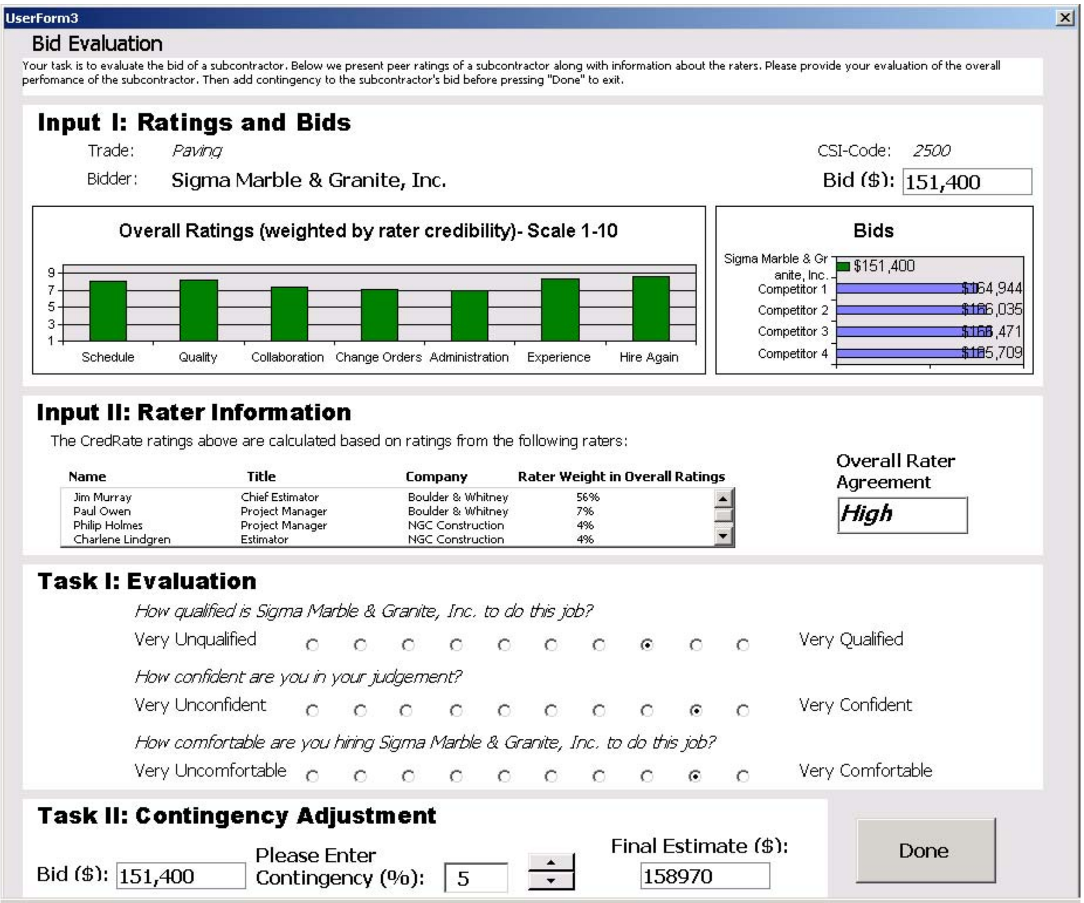
The overall rating (R_{im}) of a subcontractor m from the user i's perspective will equal the ratings provided by each rater (j) multiplied by i's estimate of j's credibility. The result is the following straightforward formula:

$$R_{im} = \sum_j R_{jm} * C_{ij} / \sum_j C_{ij} \qquad (11)$$

$$R_{jm} \neq 0$$

4.2.4.4 Step 3: Display Ratings and Rater Information

Similar to TrustBuilder I, TrustBuilder II also displays ratings and rater information. Figure 8 shows an example of the TrustBuilder II user interface, which provides the overall ratings for one subcontractor. As in TrustBuilder I, the user can see rater agreement along with total rater credibility, and input contingency for each bid.



Bid Evaluation

Your task is to evaluate the bid of a subcontractor. Below we present peer ratings of a subcontractor along with information about the raters. Please provide your evaluation of the overall perfomance of the subcontractor. Then add contingency to the subcontractor's bid before pressing "Done" to exit.

Input I: Ratings and Bids

Trade: *Paving* CSI-Code: *2500*

Bidder: Sigma Marble & Granite, Inc. Bid ($): 151,400

Input II: Rater Information

The CredRate ratings above are calculated based on ratings from the following raters:

Name	Title	Company	Rater Weight in Overall Ratings
Jim Murray	Chief Estimator	Boulder & Whitney	56%
Paul Owen	Project Manager	Boulder & Whitney	7%
Philip Holmes	Project Manager	NGC Construction	4%
Charlene Lindgren	Estimator	NGC Construction	4%

Overall Rater Agreement: ***High***

Task I: Evaluation

How qualified is Sigma Marble & Granite, Inc. to do this job?

Very Unqualified ○ ○ ○ ○ ○ ○ ○ ● ○ ○ Very Qualified

How confident are you in your judgement?

Very Unconfident ○ ○ ○ ○ ○ ○ ○ ○ ● ○ Very Confident

How comfortable are you hiring Sigma Marble & Granite, Inc. to do this job?

Very Uncomfortable ○ ○ ○ ○ ○ ○ ○ ○ ● ○ Very Comfortable

Task II: Contingency Adjustment

Bid ($): 151,400 Please Enter Contingency (%): 5 Final Estimate ($): 158970

Figure 8: User interface showing bids from and ratings of two subcontractors. The user enters contingency (risk buffer) for each bid.

The main difference from TrustBuilder I is that TrustBuilder II displays ratings for seven different criteria instead of one. The criteria are subjective measures provided by peer industry practitioners (see Section 3.1.2 of the Practical Point of Departure). To enter the ratings, the raters evaluate subcontractor performance by indicating on ten point Likert scales, the extent to which they agree with the following statements:

- **Schedule**: SubA is able to maintain schedule.
- **Quality**: SubA delivers first-class work.
- **Collaboration**: The people at SubA are responsive to other project participants regarding the resolution of any unforeseen issues.
- **Change Orders**: SubA is a “change order artist.”
- **Administration**: SubA takes care of paperwork in a fast and efficient manner.

- **Experience**: SubA is an experienced subcontractor when it comes to 05500 Metal Fabrications (Tube and Ornamental) work.
- **Hire Again**: I would be willing to hire SubA to work for me again.

Finally, Trustbuilder II also differs from TrustBuilder I in terms of the information about raters and bids that the tool displays.

- **Rater Information**: The TrustBuilder II tool shows the identity of each rater along with his/her weight in the overall ratings. The user sees the weight as a percentage measure, which indicates the weight (or credibility) of a rater relative to those of the other raters of the subcontractor. The results of Experiment I showed that the symbolic scale of TrustBuilder I could, in some cases, be confusing as well as difficult to calibrate.
- **Bid Information**: The tool shows the subcontractor's bid along with the bids from a set of competing subcontractors.

4.2.5 Comparison of TrustBuilder I and II

As the above discussion shows, the two versions of the TrustBuilder rating tool were reasonably similar. Table 17 shows the four major differences between the two tools.

Table 17 Major differences between the TrustBuilder I, and II rating tools.

Difference	**TrustBuilder I**	**TrustBuilder II**
Factors used to estimate credibility	Organizational Credibility (Co) Personal credibility (Cp)	Know Rater (KR) Same Organization (SO) Rater Expertise (X) Rater Trustworthiness (TW)
Formula to estimate credibility	Function similar to logistic regression **Error! Objects cannot be created from editing field codes.**	Exponential Function $C_{ij} = \exp(\alpha + \beta_{KR} KR_{ij} + \beta_{SO} SO_{ij} + \beta_X X_{ij} + \beta twTW_{ij})$
Information displayed about rater credibility	Symbolic Scale	Percentage measure showing relative weight of rater
Number of criteria Rated	1	7

4.3 Experimentation

In this research project I have used experimentation as the primary research methodology to investigate and validate the impact of source credibility in AEC rating systems. The first part of this section contains a discussion of alternative research methodologies. I conclude that, for this research project, experimentation is the most applicable approach. The subsection discussion covers critical issues in the experimental design, and the chapter finishes by presenting the fundamental research hypotheses, which the experiments investigated.

4.3.1 Alternative Research Methodologies

There are several research methodologies that we can apply to investigate the added value of source credibility in the context of AEC-Bidding. Below I discuss the pros and cons of four important alternative approaches.

4.3.1.1 Simulation

Simulation has been used to investigate the performance of rating systems. Researchers [65] have, for example, simulated different user behavior to compare the output of a proposed rating mechanism to that of Amazon's standard rating system. The problem with simulation is that it requires assumptions about user behavior. Given a set of user behaviors, researchers, can deploy rating mechanisms to simulate the evolution of the ratings and determine if the output is realistic. Since in Trustbuilder, the weights depend on the users beliefs at any given point in time, simulation becomes difficult. As a result, a simulation involving TrustBuilder would require a significant number of assumptions about user behaviors such as: If a user is providing false ratings; 1) what are the chances of getting detected by other users; and 2) how would these users modify their evaluations of rater credibility?

In my opinion it is far from clear how to make these assumptions which creates the additional difficulty of validating the simulation model. Simulation would be more applicable if rating systems were common in AEC practice, which would enable the validation of assumptions about user behavior through observation.

4.3.1.2 Survey

A paper or web-based survey among industry practitioners could investigate the applicability of rating systems in AEC. The advantage of this methodology is that we can administer a survey to a large number of potential users, which would facilitate statistical analysis. However, there are several problems associated with this methodology. First of all, it would be difficult to ensure a large number of responses from the desired subjects. Secondly, the evaluations would take place in a non-controlled environment, making it hard to warrant the quality of the responses. Finally, and most importantly, the context of a survey is by nature less rich than that of, for example, an experiment. As a result, even though one could assess the users' attitudes towards different rating tools, it would be practically impossible to compare the rating tools' performance.

4.3.1.3 Intervention Study

Thomsen et al [162] presents prospective validation, or intervention study, as an applicable validation methodology for simulation models, which support managerial decisions. Prospective validation consists of having industry practitioners apply the tool in a real industry setting. Industry practitioners use the research tool to detect problems or opportunities. A decision-maker who acts based on the information provided by the tool and, as a consequence, ameliorates the final outcome, provides validation of the tool's value from a managerial perspective. The problem of using prospective validation in this research project is that rating systems are currently not used in the industry. Decision-makers would therefore be reluctant to let a new, untested rating tool influence a critical task such as bidding. In addition, it would be hard to identify the cause of any differences in performance. An improvement could be due to that the tested tool was based on source credibility, or simply due to that any rating tool is better than the existing, paper based, process.

4.3.1.4 Experimentation

Experimental design allows the researcher to precisely specify and manipulate the source or message characteristics, which she is comparing [163]. In the field of source credibility theory, several researchers have used experimentation to investigate the constructs applicability in different commercial settings ([26, 28, 146].) In AEC research, a commonly conducted type of experiment is the Charrette method. Charettes are designed to evaluate the usefulness of a software application in a realistic setting [164]. In a typical Charrette, the researchers gather a group of industry practitioners who accomplish the same task in two different ways: using a traditional method, and applying an new software application. The advantage of Charrettes is that they provide a controlled environment, which allows for the completion of more complex tasks than just completing a survey. The problem is that there are practical limits to the number of participants in the Charrette, which makes statistical analysis difficult.

The experiments, which this research project involves, are different from a typical Charrette. First of all, the tools and the experiments were integrated. The users followed the instructions of a software program, which ran the experiment and also calculated the ratings. In the experiments, the users evaluated a set of subcontractors using two different rating tools. The result was a more controlled environment than the typical Charrette, which facilitated the investigation the research question in terms of its two dimensions: operationalization, and added value. Another difference from the typical Charette was that the experiments took place one by one at the participant's place of work. This approach served to minimize the time the users had to spend in order to participate, and therefore this method facilitated the recruitment of relatively large sets of testers. A controlled environment, along with a large set of users, increases the chances of obtaining statistically significant results.

A potential pitfall of the experimentation is that the controlled environment makes the context less rich, and therefore also potentially less realistic, compared to, for example, an intervention study. I argue that, for the purpose of this research project, the need to isolate the impact of source credibility requires the controlled environment of an experiment. I also argue that, in the context of AEC bidding, it is possible to construct a controlled, yet realistic, experimental validation environment. In this research project, I have therefore chosen to use experimentation as the primary means of investigation.

4.3.2 Choice of Benchmark rating tool

One way to test decision support tools is to let users apply different types of tools to solve the same problem (e.g., [165]). Researchers can then study if different methods generate similar results, and draw conclusions about their performance. Before evaluating AEC rating tools such as TrustBuilder II in this manner, it is necessary to choose a benchmark-rating tool to use as a comparison. There are two obvious candidate benchmark-rating tools:

4.3.2.1 Existing Practice (No Tool)

Today, most general contractors do not use any computerized rating tools to evaluate subcontractors. One option is therefore to compare the performance of a prototype-rating tool to the existing paper-based practice. Clayton et al [164] did, for example, in the of context design evaluation, conduct a Charette test which compared the performance (in terms of speed, accuracy, reliability and learning) of a software tool and the conventional manual process. Moreover, Kim et al [83] used this approach to validate the usefulness of an agent-based subcontractor scheduling software. However, a computerized rating application does not only help the user to aggregate information, but also to search for and retrieve the ratings submitted by different raters. While a credibility-weighted rating mechanism is unique in the way it aggregates ratings, any rating application can search and retrieve the different ratings. If the participants are subject to sufficient time pressure and information overload, we would expect any rating tool to perform better than the conventional process. Therefore, if a rating tool based on source credibility weighted rating tool did perform better than a manual rating/evaluation process, it would be difficult to identify the real cause of the improvement. As a result, it would be difficult to demonstrate the added value of source credibility in an AEC rating tool if we used the manual process as a benchmark measure.

4.3.2.2 Unweighted Rating Tool

Another potential benchmark measure is an unweighted tool, in which all the ratings weigh the same. Unweighted rating tools, which are currently found in consumer e-commerce (Bizrate), and which are planned for AEC e-commerce (RatingSource), have the advantage of being simple to deploy and use. In the near future, rating tools of this type are likely to be the most common in industry practice. I have therefore chosen to use an unweighted tool as the benchmark measure in this research project. As a result, I can demonstrate whether, the added complexity of a credibility weighted rating tool, adds any value for a user who was previously using a more common and simpler rating model. However, I should note that comparing the performance of a credibility-weighted tool to an unweighted tool will not demonstrate that the credibility-weighted tool is the best tool in an absolute sense. The point of departure chapter of this dissertation presents a number of potential solutions, and, in theory, we could compare the performance of a credibility-weight rating system to all of these tools. However, I claim that implementing each of these solutions is beyond the scope of this project. Another alternative is to implement just one of these alternative solutions, but since none of them have been established in practice, it would be difficult to determine which one to use as a benchmark. Furthermore, the calibration of, for example, a collaborative filtering mechanism would require a substantial amount of clean data, which would be very difficult to obtain in the AEC industry, where ratings are seldom recorded.

4.3.3 Overview of the experimental design of two experiments investigating source credibility in AEC e-bidding

4.3.3.1 Introduction

To investigate the added value of source credibility in the context of AEC e-bidding, I conducted two experiments (Experiments I and II) where users deployed a prototype rating tool to evaluate subcontractors bidding for jobs from a general contractor. In this section I discuss the experimental design of the two experiments while the Results chapter (0) provides a more detailed description of the two experiments.

In experimentation, a researcher can control the conditions to limit the influence of factors other than the independent variables. If the fundamental research hypothesis is valid, this approach substantially increases the chances of obtaining statistically significant results. However, by controlling the environment a researcher runs the risk of making the task unrealistic to the users. In the case of a rating application, users would most likely object if they were asked to assess thirty different painting contractors whose bids were exactly 8.2% lower than the second lowest bid. If, on the other hand, the environment is made more realistic by allowing external factors (such as trade and bid size) to vary, it becomes more difficult to draw conclusions about the impact of the independent variable (the type of rating tool) on the dependent variables. For instance, it can be difficult to determine if the variation in risk buffer is caused by the bid size or the type of rating tool.

Recognizing the need for the experimental environment to be realistic as well as controlled, I conducted two separate experiments. The environment of Experiment I was very controlled, while Experiment II involved a less controlled, and therefore also more realistic environment. I argue that consistent results across two experiments which; 1) test two different ways of operatationalizing source credibility theory (TrustBuilder I and II); 2) involve two different groups of users (experts and non experts); and 3) provide two different kinds of environments, in terms of control and realism; increases the generality of the results. I next discuss the main differences between the two experiments, in terms of participants, subcontractor ratings, participants, credibility weighted tool, and bid variation.

4.3.3.2 Participants

The participants in Experiment all had construction management experience, but they were non-experts in subcontractor evaluation. The user group was also heterogeneous, since it consisted of AEC students and faculty, as well as industry practitioners from three different continents. One major advantage of conducting the first experiment with the non-experts was that it enabled the refinement of the rating tool before testing it with industry practitioners. Having validated the underlying research hypotheses, I performed the second experiment with participants who were highly qualified AEC practitioners, very familiar with the task of evaluating subcontractors.

4.3.3.3 Subcontractor Ratings

To isolate the impact of source credibility, Experiment I involved hypothetical subcontractor ratings. In addition, Experiment I created a unique set of raters for each user. First the user provided the names of three persons whom he or she knew at three different AEC companies. The user then assessed the credibility of these three persons, three unknown project manager working for the same companies, as well as an unknown project manager working for an unknown contractor. This approach allowed TrustBuilder I (the tool used in Experiment I) to estimate the credibility of a set of raters, some of whom the user knew, and others who were unknown. The tool then attributed hypothetical subcontractor ratings to the different raters. Finally, the participants used these ratings evaluate a set of subcontractors.

Experiment II, on the other hand, was designed to be as realistic as possible, and hence involved actual ratings of subcontractors. The inclusion of real subcontractor ratings created two additional requirements, regarding the set of the industry participants who took part in Experiment II. First of all, the participants had to rate the subcontractors before the experiment took place. Moreover, the participants all had to belong to the same geographical community to ensure that each person knew at least some of the people who had rated the subcontractors.

4.3.3.4 Credibility Weighted Rating Tool

In Experiment I the users tested the basic TrustBuilder I tool, while Experiment II involved the more refined TrustBuilder II tool. The major differences between the two tools lie in the algorithm used to estimate credibility, and the information available to the user. In the context of experimentation, the advantage of the TrustBuilder I tool is that the user has less information to work with, since it only shows the ratings for one overall criterion. This approach makes it easier to isolate the impact of source credibility on user behavior. In the second version the participants used the more advanced TrustBuilder II tool, in which the user can see the ratings on seven separate criteria. Providing more information to serve as a basis for bid evaluations increases the realism of the task from the point of view of the expert participants in Experiment II.

4.3.3.5 Bid Variation

Another difference between the two experiments is the variation of the bid amounts. Bid amounts could potentially influence user behavior in two ways. First of all, if the subcontractor's bid is substantially lower than competing subcontractors, a user may suspect that the subcontractor has left out some critical element when estimating the cost of the job, and will therefore add some extra contingency. Secondly, the risk that the general contractor is exposed depends on the size of the subcontract. If the total cost of a project is $3M; a 10% cost overrun on a $5,000 painting subcontract will not impact total profits; while a 10% overrun on a $500,000 HVAC job will certainly impact the bottom line. In order to isolate the impact of source credibility in Experiment I, I chose to minimize the variation of bid amounts. As a consequence, for each trade, the user interface only showed two competing subcontractors. The two bids only varied by approximately 5% around a baseline bid of $16,000.

In Experiment II, on the other hand, the bid amounts reflected those of the original project and therefore ranged from $5,000 to $400,000. The user interface showed a low bidding subcontractor along with bids from five competitors. The distance between the low bid and competing bids varied randomly, but was, on average, 15% of the low bid. The 15% variation was based on the variation of the bids on the test case project in Experiment II, along with discussions with the project manager.

4.3.3.6 Summary of differences in the design of Experiment I and II

Table 18 summarizes the major differences between Experiment I and Experiment II in terms of research design:

Table 18 Comparison of Experiment I and 2. The key differences between the two experiments relate to the participants, the type of ratings supporting subcontractor evaluation, and the version of the TrustBuilder rating tool that the participants were using

	Experiment I	**Experiment II**
Participants	*Non-experts* 16 AEC students, faculty and practitioners. All familiar with but not specialized in evaluating subcontractors	*Experts* 15 Bay Area project managers, estimators, and operation managers, specialized in evaluating subcontractors.
Subcontractor Ratings	*Hypothetical* Ratings of one overall criteria	*Real* Ratings by peer industry practitioners of 26 Bay Area subcontractors of seven subjective criteria
Version of TrustBuilder credibility-weighted tool used to evaluate subcontractors	Basic TrustBuilder I	*Refined* TrustBuilder II
Bid Variation	*Low* UI shows two competing subcontractors whose bids vary by approximately 5% around a baseline bid of $16,000.	*High* UI shows Low bidder along with 5 competing bidders. Bids vary by, on average, 15%.

4.3.4 Fundamental Research Hypotheses

The two experiments investigated the research question through a set of research hypotheses. It is important to note that a rating system is similar to an expert system, and could therefore be expected to fulfill the same requirements in terms of validation. In the context of expert systems "validation is a general term that embraces verification (i.e., testing that an implemented program meets its specification), but also is concerned with accuracy of advice, quality of recommendation, soundness of underlying model, and even user acceptability [166].” This definition is similar to my division of the research question into two sub-questions, which relate to operationalization and added value (See Problem Chapter (Section 2.3.) In this section I will propose a set of fundamental research hypothesis, which address each of the two sub-questions. These fundamental research hypotheses are fairly abstract. In the presentations of Experiments I and II in the results chapter, I state a set of detailed hypotheses, which are specific versions of the fundamental hypotheses, as they are applied in each experiment.

4.3.4.1 Research Hypotheses referring to first part of research question: *How is it possible to operationalize source credibility to support the calculation of weights that are based on rater identity in an AEC rating tool?*

The first sub-question, relating to operationalization, corresponds Wright and Bolger’s [166] definition of the verification part of the validation of an expert system. The best measure of to what extent the operationalization of source credibility is successful, is the extent to which it predicts the rater weights that users assign to different raters. The pair-wise comparison module of the two tools measures the user’s estimate of rater weights. Both experiments can therefore investigate the following fundamental research hypothesis:

A credibility weighted model will better model the rater weights expressed by users in pair-wise comparisons than an unweighted model.

Another feature of good model for rater credibility is that it only contains factors, which influence rater weight. In other words, the model should be as small as possible but still contain the main factors, which significantly contribute to the estimation of rater weight. As a result, a second hypothesis relating to operationalization is:

The factors used in the credibility weighted model all influence rater weight.

4.3.4.2 Research Hypotheses referring to second part of research question: *How can a rating system based on source credibility theory add value in the process of evaluating AEC subcontractors?*

For a decision support tool to add value, it should change user decisions. In an e-AEC rating system the key decision is the users' evaluation of expected subcontractor performance. To what extent does the user let the rating system influence her evaluations of subcontractors? Can the user trust the output of a rating system enough, to let it influence her evaluations of subcontractor performance? I therefore state the following hypothesis:

Users will vary their subcontractor evaluations more when using the credibility weighted rating tool than when using the unweighted rating tool.

Another measure of trust in ratings is the confidence that the user expresses in her evaluations of subcontractors based on information provided by a rating tool. A reasonable interpretation is that the more confident a user is in her evaluations, the more added value the tool provides. The corresponding research hypothesis is the following:

The use of a credibility-weighted tool instead of an unweighted tool will positively affect the users' confidence in the accuracy of their judgments.

Finally, the perceived usefulness of a rating tool is another measure of added value. The associated hypothesis is as follows:

Users estimate that a credibility-weighted tool would be more useful to use AEC subcontractors than an unweighted tool.

5 Results

5.1 Introduction

This chapter presents and analyzes the results from two experiments. The two experiments are referred to throughout this document as *Experiment I* and *II*. As I stated in the Problem Chapter (1), the fundamental research question of this project is as follows: *How can source credibility theory support rating systems in the procurement of AEC services?* The two experiments investigate the research question in terms of the two sub-questions: 1) *How is it possible to operationalize source credibility to support the calculation of weights that are based on rater identity in an AEC rating tool? ;* and 2) *How can a rating system based on source credibility theory add value in the process of evaluating AEC subcontractors?*

This chapter discusses both experiments in terms of research hypotheses, method and results. The first section covers Experiment I where the users who all had construction management experience, but who were non-experts regarding the task of evaluating subcontractors, used the basic TrustBuilder I tool to evaluate a set of fictitious subcontractor bids. The chapter will then discuss Experiment II, which was designed to repeat the experiment with participants who were industry practitioners with extensive experience in evaluating bids. The participants in Experiment II used the more refined rating TrustBuilder II and based their evaluations on actual ratings of existing Bay Area subcontractors. I argue that the fact that the two experiments had similar results, even though they involved different versions of the TrustBuilder tool and used two different sets of participants, provides evidence for the generality of the results. In the last section of this chapter, I summarize and discuss the results of the two experiments.

5.2 Experiment I: 16 AEC students, faculty and practitioners testing TrustBuilder I

5.2.1 Introduction

This section discusses the design and results of Experiment I, the first of two experiments that investigated the applicability of source credibility theory as a basis for an AEC rating system. The experiment uses a 'within subject' design, which compares the participants' behavior and attitudes when using three different tools to evaluate subcontractors:

Unweighted Ratings: The tool calculates the subcontractor's overall ratings as the average rating where all raters are weighted the same. The user can see the number but not the identity of the raters who have rated the subcontractor. This is the standard rating mechanism and is similar to that used by RatingSource and Bizrate's systems.

Credibility-weighted ratings: The TrustBuilder I tool, described in section (4.2.3), constituted the credibility-weighted tool in Experiment I. The tool weights each subcontractor rating by the user-defined credibility of the rater. When evaluating subcontractors, the user can see the overall rating along with the total credibility of, and agreement between, the raters. The major purpose of the experiment was to compare the performance of this mechanism to that of an unweighted mechanism.

No ratings: In this tool, the users do not have any ratings to support the subcontractor evaluations. This mechanism provides a baseline measure to which I could compare the two rating mechanisms.

The participants used the three tools to evaluate subcontractors bidding for the trades subcontracted in the construction of a recently completed $5M office building in San Francisco, California. The purpose of the experiment was to investigate whether taking into account rater credibility would aid the decision-maker when he or she is evaluating bids from a previously unknown subcontractor.

5.2.2 Research Hypotheses

To investigate the research question in terms of its two dimensions, operationalization and added value, the experiment evaluated the two rating models in terms of a set of research hypotheses**.** The first three hypothesis predict that a credibility-weighted tool will be superior to an unweighted tool in terms of, ability to predict rater weight, influence on bid contingency, as well as user confidence; while the last three hypotheses predict that information displayed in the credibility weighted tool will influence bid contingency.

Given that the first goal of the experiment was to test the operationalization of the credibility-weighted model (TrustBuilder I), a first step is to ensure that that the credibility-weighted model does in fact model rater weights better than an unweighted model. If this is not the case, we could expect the effects associated with the use of a credibility-weighted rating tool to be very marginal. I therefore posed the following hypothesis:

Hypothesis 1: In the context of participants making pair-wise comparisons, credibility measures calculated using the TrustBuilder I methodology are better than an unweighted (constant) model at predicting the relative weights users attribute to different raters.

Next, the experiment investigated the added value provided by a source credibility weighted tool in AEC e-bidding. The first research hypothesis in this section refers to user behavior. The underlying assumption is that the more a user trusts the ratings of a rating tool, the more she will let them affect her decisions. In this case, the user's decision involves the assignment of the bid contingency, or risk buffer, added to the subcontractors' bids. Since bid contingency is the only means by which the user evaluates the subcontractors' performance, the decision also represents her estimation of overall subcontractor performance. The resulting hypothesis is*:*

Hypothesis 2: Users will vary the contingency added to the subcontractors' bids more when using the credibility weighted tool than when using the unweighted tool.

A second measure of user trust was the user's confidence in his or her judgment when evaluating overall subcontractor performances. This is a direct attitudinal measure and the associated hypothesis is:

Hypothesis 3: Users are more confident when using a credibility-weighted rating tool than when using an unweighted rating tool.

The experiment also investigated the influence on user behavior of three separate measures providing information about the subcontractor and the raters. The TrustBuilder I bid evaluation interface provides the three measures, which are Aggregated Rating (overall rating of the contractor weighted by rater credibility), Rater Agreement, and Total Rater Credibility. The purpose was to investigate whether they do in fact affect the contingency added to the subcontractor bids. The experiment investigated three hypotheses (Hypotheses 4-6), which predict that the aggregated ratings, rater agreement, and total credibility are negatively correlated to the risk buffer added to the bids. The three hypotheses are as follows:

Hypothesis 4: The aggregated rating of a subcontractor is negatively correlated to the contingency added to the subcontractor's bid.

Hypothesis 5: The rater agreement regarding the performance of a subcontractor is negatively correlated to the contingency added to the subcontractor's bid.

Hypothesis 6: The total credibility of all the raters that have rated a subcontractor is negatively correlated to the contingency added to the subcontractor's bid.

5.2.3 Method

5.2.3.1 Participants

In Experiment I the participants consisted of sixteen construction management students, faculty and professionals with ages ranging from 24 to 55 (M= 34.5, SD=9.3.) They were all familiar with AEC bidding and fluent in English, even though they were of various origins (European= 8, Asia=6 and North America=2). I randomly assigned the participants to the order in which the different rating tools were presented.

5.2.3.2 Procedure

Experiment I was a within subject design with the type of tool as the within subject factor. It was carried out on an individual basis with each participant supervised by an instructor. The experiment took place on a personal computer at the participant's place of work. The instructor began by showing the participant a ten-minute presentation, which introduced the concept of rating systems in AEC, before pressing "Run" to start an application which uses the Microsoft Excel interface.

At the start of the application, participants read an introduction which informs them that they are going to act as a project manager for a general contractor; that their task is to compose a bid for the construction of a small office building; and that they will have three different Internet-based rating systems to assist them. The participants were then asked to name three different people in the construction industry working for separate organizations, before using the multi-dimensional McCroskey source credibility scale to rate seven different people. The rated people comprised the three persons the participants had named, three unknown project managers working for the same companies as the three named persons, and an unknown project manager working for an unknown contractor. Next the participants calibrated the rating tool by making pair-wise comparisons of the credibility of the different raters as described in Section 4.2.3.3. The participants evaluated divergent ratings from all possible (twenty one) pairs of raters before the system performed a logistic regression. Participants then proceeded to the next task, which consisted of using one of the three rating tools to evaluate a pair of bidding subcontractors for each of the job's seventeen trades. Each tool showed the participants bids from the two subcontractors and asked the participants to add bid contingency along with selecting the best bidder. In the next step, the participants were asked to provide information for the general contractor's "risk management program." The users assessed the likely performance (along with the confidence in their judgment) of a pair of bidders rated shown in each of the three tools. Finally, the participants filled out a final questionnaire about their attitude towards rating systems and their experiences during the experiment.

5.2.3.3 Manipulation

Given the exploratory nature of the study, the experiment was carried out using hypothetical data. For each of the seventeen trades, I created two competing subcontractors. Data were also generated four eight raters of the thirty-four subcontractors. The values of the subcontractor rating were random but distributed in such a way that each rater had evaluated some but not all of the subcontractors. The first three of the eight raters were given the identity of three persons known to the participants; the next three were given unknown names but assigned to the same organizations as the first three; and the last two were completely unknown to the participants. The subcontractors also bid random amounts, but for each trade the two bids were correlated so that they did not differ by more than 5%. The tool provided three different user interfaces corresponding to the rating systems (unweighted, credibility weighted, and no ratings). Figure 1 in Chapter 4 shows the user interface for the credibility-weighted system, which was the one implemented in TrustBuilder I. The unweighted rating tool was a simplified version of the credibility-weighted rating tool. The user could see the average rating on both a symbolic and continuous scale along with the number of people who had rated the subcontractor. However, she did not know who had rated the subcontractor. The tool without ratings was very simple. It consisted of the two subcontractors' names and bids.

5.2.3.4 Measures

Rater credibility was measured with the McCroskey 12-item credibility scale.

Goodness of fit of model predicting rater weights was measured using the sum of squared errors in the pair-wise comparisons for the two models.

Bid contingency was measured with a single item. The users entered a number between 0-100% for bid contingency. This number was intended to reflect the participants' assessment of the risk buffer that should be added to the bid as well as the extra cost of managing an under-performing subcontractor.

Users' confidence in their assessments was measured with a single item question: "How confident are you in your judgment?"

Risk Attitude: The experiment measured the user's risk attitude by letting them make pair-wise assessments of two pairs of unknown subcontractors. Participants who put on average more than 55% weight on the negative rating were classified as risk averse.

5.2.4 Results

5.2.4.1 Operationalization

To test whether the credibility-weighted model was better than the unweighted model at predicting rater weights (Hypothesis 1), I calculated the total error of the 336 pair-wise comparisons provided by the sixteen users. The credibility function (i.e., Equation 6) was estimated at the individual level using the twenty-one comparisons that each user provided. Figure 9 shows that the total squared error was much higher for the unweighted model (17.52) than for the credibility-weighted tool (5.11). To investigate the significance of this difference, I performed a generalized maximum likelihood ratio test. This type of test usually "performs reasonably well" in situations in which the hypotheses are "not simple" and, as result, for which there exist no standard optimal test (e.g., t-test, F-test) [149]. In this case, the maximum likelihood ratio test enabled a comparison between the unweighted and the credibility-weighted models in terms of the sum of the squared errors, taking into account the higher degrees of freedom (48 vs. 0) of the credibility-weighted model. The maximum likelihood ratio test showed the difference to be significant ($p < .001$).

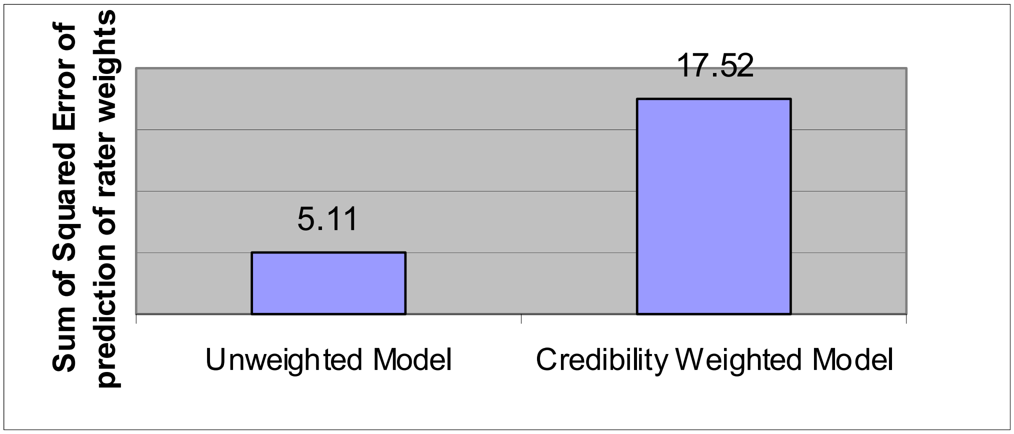


Figure 9: When predicting the users' assignments of rater weights the sum of squared errors were considerable smaller in the credibility-weighted model (5.11) than in the unweighted model (17.52) (Maximum Likelihood Ratio Test, N=336, p< .001).

The difference in errors indicates that the credibility-weighted rating model is superior to the unweighted model when it comes to predicting the weight of raters and, therefore, provides evidence that the proposed operationalization of source credibility theory can be used to model weights in the context of AEC bidding. This is a prerequisite for the application of a successful rating tool.

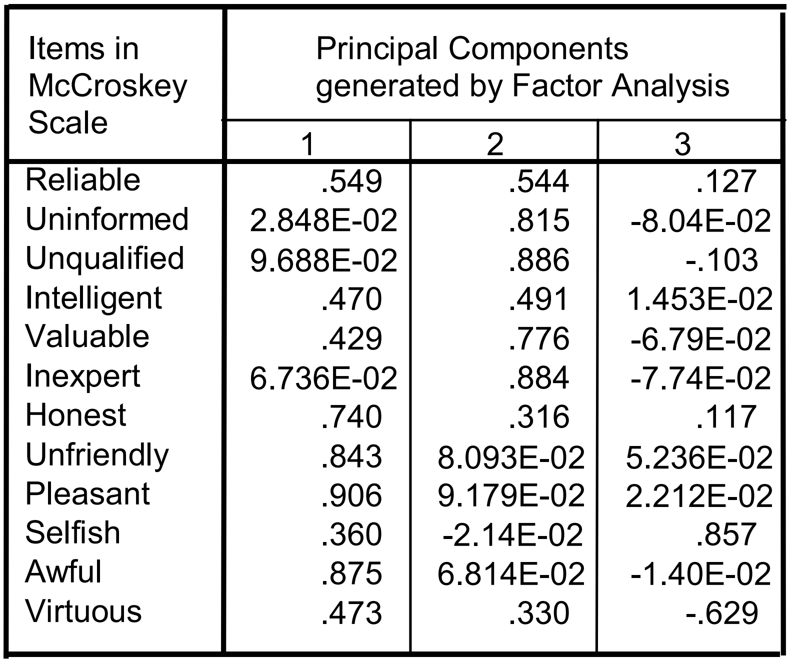
To further test the applicability of source credibility in an AEC bidding context, I also performed a principal component factor analysis. A principal component analysis is an eigenanalysis technique, which extracts a set of eigenvectors and their associated eigen-values by a step-wise procedure, where each eigenvector, or component, accounts for a maximum amount of variance. The purpose of the factor analysis was to investigate whether the McCroskey scale items did factor into the two components: "perceived expertise" and "perceived trustworthiness." I limited this analysis to those cases where participants knew the raters personally. Figure 10 displays the results, which were encouraging but non-conclusive. As shown in Figure 10, the factor analysis[7], which uses cut-off eigen value of 1, resulted in three components. The first two corresponded to the expected outcomes, perceived trustworthiness and expertise, respectively. However, two scale items expected to belong to the expertise factor (reliability and intelligence) were

[7] The factor analysis used varimax rotation with Kaiser normalization. The rotation converged in 4 iterations.

evenly distributed between the two factors. Given its dual meaning in English,[8] the word "reliable" was expected to cause some problems but the even distribution of "intelligent" was more surprising. One explanation may be the sensitive nature of this question among the European participants, two of whom pointed out that they did not feel comfortable rating their peers' intelligence. A final unexpected outcome of the factor analysis was that it generated selfishness as a third separate factor. Still, the results indicate that, measuring the credibility in terms of the two components Expertise and Trustworthiness seems to be valid in AEC bidding as well. As a result, it would be appropriate to measure Trustworthiness and Expertise as separate factors with individual weights in the logistic regression models of rater credibility. Another way to capitalize on this result would be to reduce the number of items measuring each factor (expertise and trustworthiness) in order to save time for busy users rating the credibility of people and organizations in the industry.

However, even though the credibility-weighted model seemed to perform better than the unweighted model, it was not problem-free. Finding the accurate zero point for the factors C_p and $C_{o,}$ which model personal and organizational credibility (see section 4.2.3.2), turned out to be one problem area in the calibration phase. Some users gave a known rater lower credibility ratings than a rater for whom they knew only the organization, but still found the former more credible when making a pair-wise comparison. When later asked about the reasons for this behavior, they explained that even though they perceived the known person as not being very credible it was "someone [they] know". This result indicates that a more elaborate model than simply subtracting by the appropriate C_o and C_u is called for. I believe that most problems can be addressed by refining both the operationalization of the McCroskey and the user-interface design.

[8] The dual meaning of "reliable" is illustrated by the fact that, for example, the Microsoft Word Thesaurus gives "trustworthy" as a synonym for "reliable". In the McCroskey scale, on the other hand, "reliable" belongs to the Authoritativeness (Expertise) factor.



Items in McCroskey Scale	Principal Components generated by Factor Analysis		
	1	2	3
Reliable	.549	.544	.127
Uninformed	2.848E-02	.815	-8.04E-02
Unqualified	9.688E-02	.886	-.103
Intelligent	.470	.491	1.453E-02
Valuable	.429	.776	-6.79E-02
Inexpert	6.736E-02	.884	-7.74E-02
Honest	.740	.316	.117
Unfriendly	.843	8.093E-02	5.236E-02
Pleasant	.906	9.179E-02	2.212E-02
Selfish	.360	-2.14E-02	.857
Awful	.875	6.814E-02	-1.40E-02
Virtuous	.473	.330	-.629

Figure 10 Output from factor analysis of the items measured by McCroskey Source Credibility Scale. The factor analysis generated the two expected components Authorativeness (Expertise) and Character (Trustworthiness). Less expected was that Intelligence and Reliability would distribute evenly across the two components and that selfishness would be a third independent factor.

5.2.4.2 Risk Attitude

The Experiment also investigated the participants' risk attitude. As section 3.2.6.2 in the point of departure shows, the framing of a message, in combination with a user's risk averseness, can interact with the source's perceived credibility when determining the weight a user attributes to information. The results showed that four out of the sixteen users turned out to be substantially risk averse. The remaining twelve turned out to be risk neutral or only marginally risk averse. One interpretation of this finding is that it supports a customization of the tool that could incorporate each user's risk attitude. On the other hand, it also shows that risk averseness is not a major factor in the pair-wise comparisons. We would expect risk adverse users to find ratings, which are weighted by rater credibility, to be less useful. If a subcontractor receives low ratings, this outcome would be very important to a risk adverse user, even in the case she perceives the rater's credibility to be very low.

5.2.4.3 Variation of Bid Contingency

Experiment I investigated the participants' evaluations of subcontractor performance by calculating the variance of the contingency added to the bids. Figure 11 shows the users' mean variance using the three different rating tools. The result, that the credibility-weighted tool produces greater variance (M=8.82) than the unweighted tool (M=3.56), which in turn produces greater variance than no tool (M=0.81), is consistent with Hypothesis 2. When the users have no tool, and thus no information about subcontractor performance, their natural strategy is to add a constant contingency to each bid, which results in no variance. Using the unweighted tool, the user has information that is superior to no ratings, and therefore varies the bid contingency more. Finally, using the credibility-weighted tool, the users, confident that the ratings from the credible source will be given higher weight, vary their evaluations even more.

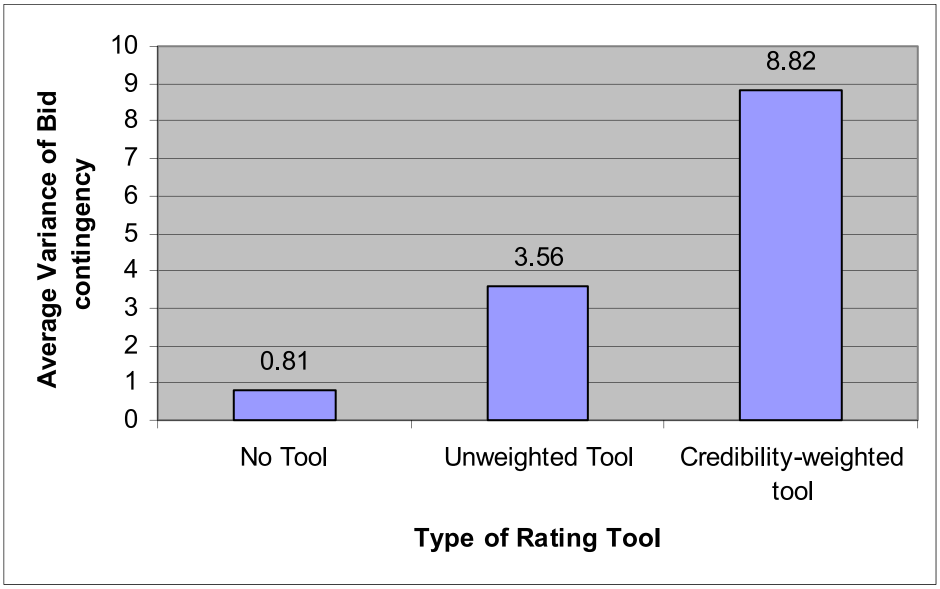


Figure 11 Average Variance of Bid Contingency using the three different tools. The users had a much higher variance using the credibility-weighted tool, which indicates that they let this tool influence their decisions to a greater degree.

A Wilcoxon matched-pairs signed-ranks test evaluated the statistical significance of the above results. This test ranks the absolute differences for all the matched pairs before summing up the negative and the positive ranks. As a result, The Wilcoxon test, which is weaker than a t-test, does not require the assumption that the variables are normally distributed. In this case, each matched pair corresponds to each user's variance when using two different tools. Following the criteria and procedures set out in Cohen and Holliday [167] I deemed the form of data yielded suitable for analysis by the Wilcoxon test and calculated the variance of the contingency assigned by each user for each tool (Figure 12). The Wilcoxon matched-pairs signed-ranks test (W+ = 10, W- = 126, N = 16, p < .0014) showed that users varied the contingency to a greater degree using the credibility-weighted tool (Hypothesis 2). Similarly, the difference in variance was significant when comparing the credibility weighted to the use of no tool (Wilcoxon matched-pair signed-ranks test: W+ = 133, W- = 3, N = 16, p < .00016).

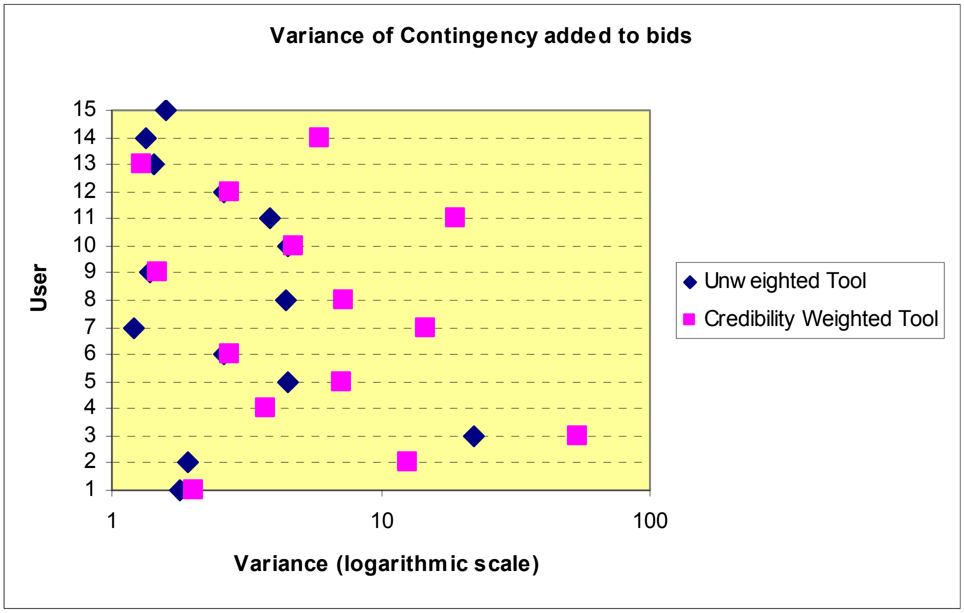


Figure 12 Variance of contingency for each user with the unweighted and credibility weighted tools. Fourteen out of sixteen users had a higher variance when using the credibility-weighted tool (Wilcoxon: p<.005).

The results indicate that the users will vary their decisions about the evaluation of subcontractors to a greater degree when using a credibility-weighted tool. As a result, the bidding price will be of less importance, and a user would be less likely to select the lowest bidder than he or she would when using the unweighted tool. This is an important result since the purpose of a decision aid tool such as a rating system is to provide the user with information that she trusts enough to act upon.

5.2.4.4 Confidence

Experiment I also showed that the participants expressed higher confidence in the credibility-weighted tool compared to the unweighted tool (Hypothesis 3). A pair-wise t-test evaluated statistical significance of this result since this type of test is typically used to investigate variables that are measured using standard Likert scales [168]. Figure 13 shows that the participants expressed higher confidence in their evaluations when using the credibility weighted tool (M=5.97, SD=2.00) than in the unweighted tool (M=5.00, SD=3.83, N=16, paired t-test: p<0.005.). Similarly, the confidence was higher when using the credibility-weighted tool than when using no tool (M=3.15, SD=2.19, N=16, paired t-test: p<0.005). The results for the attitudinal confidence measure were therefore consistent with the results for the behavioral measure of bid contingency described above. Confident users will be more likely to vary their decisions depending on the information provided by the rating tool.

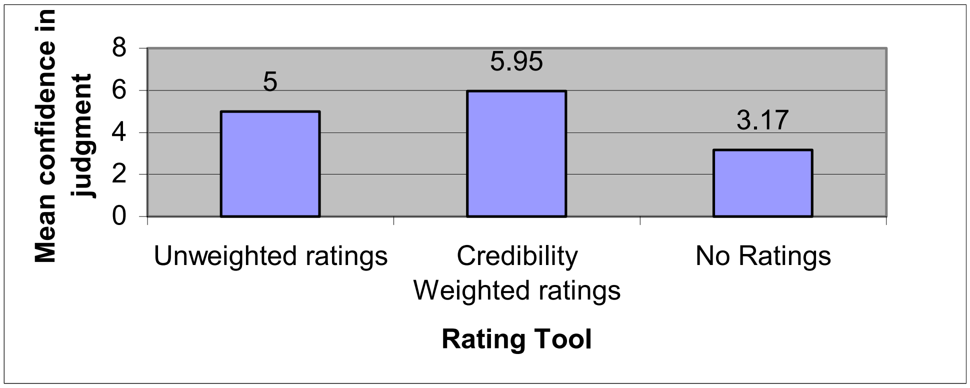


Figure 13: Confidence in judgment using three different rating tools. Users expressed more confidence when using the credibility-weighted tool than with the other two tools (t-test: p<.05) .

5.2.4.5 Agreement and Total Credibility

The study also investigated the extent to which Aggregated Rating, Agreement and Total Credibility influenced bidding decisions. A t-test showed all three measures - Aggregated Rating (p<.0001), Rater Agreement (p<.0001), and Total Rater Credibility (p<.00005) - to be significant predictors of bid contingency. As shown in Figure 14, the coefficients are all negative since the better the overall rating, the higher the consensus of the raters; and the more trustworthy these raters are perceived to be, the less the user will feel inclined to add bid contingency. However, we should not that, since each user performed ten to twelve evaluations using the credibility-weighted rating tool, there are interdependence between the observations. To partly deal with this problem, I have normalized the variables for each user by converting the values to z-scores. Nonetheless, I recognize that there will still be interdependence that is not taken account for in the regression analysis. However, given that the regression shows all three factors to be highly significant (p<.0005), I argue that the outcome would not have been different if an independent user had provided each value.

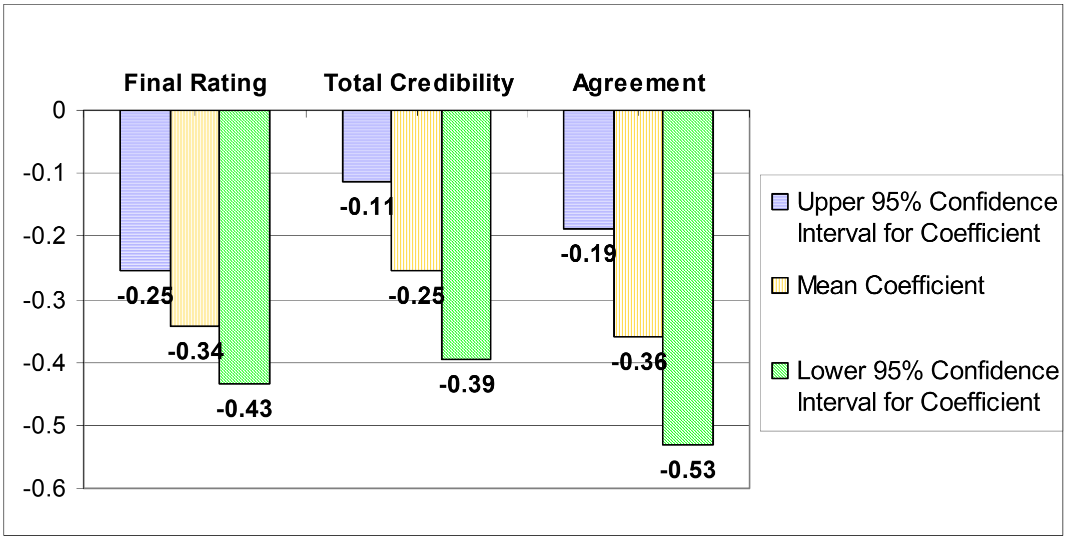


Figure 14 Coefficients in regression of bid contingency when using credibility-weighted tool. In accordance with Hypotheses 4-6 all the three measures Final Rating, Total Credibility, and Agreement, contribute to a decrease in the contingence added to a bid.

My conclusion is therefore that the results are therefore consistent with Hypothesis 4-6. Nonetheless, I should note that since the experimental design included only a limited number of raters (fewer than 8) the total credibility index did not prove to be all that useful. It was simply too easy for the users to identify the raters and assess the raters' total credibility by themselves. If the participants had been dealing with forty to eighty, rather than four to eight, raters this measure would probably have been more useful. It would then have been more difficult and time consuming for a user to assess total credibility by herself. The fact that total credibility and rater agreement do affect rater contingency suggests that such measures can be useful in a market where ratings are provided by a large number of raters.

5.3 Results Experiment II: 15 AEC practitioners testing TrustBuilder II

5.3.1 Introduction

This section describes the results of a second experiment testing the applicability of source credibility in the context of AEC ratings. The purpose of Experiment II was to repeat Experiment I in more realistic conditions. The experiment therefore involved actual ratings of AEC subcontractors and all participants were industry practitioners specializing in evaluating subcontractors. In addition, the experiment was carried out using TrustBuilder II, the refined version of an evaluation tool operationalizing source credibility (described in 4.2.4.)

The experiment was designed to compare the performance of two different rating models:

Unweighted Ratings: The tool calculates the subcontractor's overall ratings as the average rating where all raters are weighted the same. The user can see the number but not the identity of the raters who have rated the subcontractor. This is the standard rating mechanism and is similar to that used by RatingSource and Bizrate's systems.

Credibility-weighted tool: The TrustBuilder II tool, described in section 4.2.4, constituted the credibility-weighted tool in Experiment II. The tool weights each subcontractor rating by the user-defined credibility of the rater. When evaluating subcontractors, the user can see the overall rating along with the total credibility of, and agreement between, the raters.

This chapter begins by listing the research hypotheses, before describing the research methodology and presenting the results of the experiment.

5.3.2 Research Hypotheses

The experiment evaluated the two rating models in terms of a set of research hypotheses. The first three hypotheses investigate the operationalization dimension of the research question, and refer to the different models ability to predict rater weights. The remaining three hypotheses predict that a credibility-weighted tool will add more value in AEC bidding than an unweighted tool.

5.3.2.1 Hypotheses investigating model *operationalization*

As a first step, Experiment II tested if all factors in the TrustBuilder II credibility weighted model contributed to the estimation of rater weight. As a result, Hypothesis 7 relates to whether any of the factors included in the credibility-weighted model is insignificant when it comes to modeling rater weight. As described in section 4.2.4.3 Trustbuilder II estimates the rater credibility (C_{ij}) of rater j from user i's perspective as:

$$C_{ij} = \exp(\alpha + \beta_{KR} KR_{ij} + \beta_{SO} SO_{ij} + \beta_X X_{ij} + \beta tw TW_{ij}) \qquad (9)$$

Zero or negative coefficients for a factor in the model indicate that a factor is insignificant or heavily correlated with other factors.

Hypothesis 7: The factors used in the credibility weighted model influence rater weight.

More specifically:

i) *The coefficient for Know Rater (β_{KR}) in the model estimating rater weight is positive;*

ii) *The coefficient for Same Organization (β_{SO}) in the model estimating rater weight is positive;*

iii) *The coefficient for Rater Experience (β_X) in the model estimating rater weight is positive; and*

iv) *The coefficient for Rater Trustworthiness (β_{TW}) in the model estimating rater weight is positive.*

Similarly to Experiment I, Experiment II investigated whether the two models are equally good at predicting the weights in the pair-wise comparisons. These weights can be seen as the users' subjective opinion of what weights are appropriate when aggregating ratings from two different raters. I therefore again pose Hypothesis 1 of Experiment I.

Hypothesis 1: A credibility-weighted model will better model the rater weights expressed by users in pair-wise comparisons than does an unweighted model.

The final hypothesis relating to operationalization focuses on the correlation between rater credibility and the consistency between a user's and a rater's ratings. This consistency can be interpreted as an objective measure of how well a rater's ratings predict the user's own ratings. It is not at all a given that this measure will be correlated to the users' subjective evaluations of rater credibility. Still, since consistency is used to determine weight in several collaborative filtering mechanisms, an interesting hypothesis is:

Hypothesis 8: The credibility of a rater from a user's perspective is positively correlated to the consistency between the user's and the rater's ratings.

5.3.2.2 Hypotheses investigating added value of rating tool to bid decisions

The first research hypothesis in this section refers to user behavior. The underlying assumption is that the more a user trusts the ratings provided by a rating tool, the more she will let them affect her decisions. In this case, the decision is the evaluation of subcontractors and the resulting hypothesis is:

Hypothesis 9: Users will vary their evaluations of subcontractors more when using the credibility-weighted tool than when using the unweighted tool.

Experiment II tests Hypothesis 9 using two different measures of the participants' evaluation of subcontractors. Overall qualification constitutes a direct assessment of the expected performance of the subcontractor, which primarily depends on the ratings displayed. Bid contingency, on the other hand, depends not only on the user's assessment of subcontractor quality, but also on other factors such as type of trade and competing bids. Hypothesis 9 can therefore be divided into two sub hypotheses:

Hypothesis 9 I: Users will vary their overall ratings of subcontractor qualification more when using the credibility weighted rating tool than when using the unweighted rating tool.

Hypothesis 9 II (Same as Hypothesis 2 in Experiment I): Users will vary the contingency added to bids more when using the credibility-weighted than when using the unweighted rating tool.

A second measure of user trust was the users' confidence in their judgment when evaluating overall subcontractor performance. This is a direct attitudinal measure and the associated hypothesis is again Hypothesis 3 from Experiment I:

Hypothesis 3: The use of a credibility-weighted relative to an unweighted tool results in increased user confidence in the user's judgments of overall performance.

The final purpose of Experiment II was to investigate the extent to which the users found the credibility-weighted tool more useful than an unweighted tool. Originally, the credibility-weighted tool was designed to support an e-market place but interviews indicated that it could also be useful in an internal rating application. As a result, I posed the following two hypotheses:

Hypothesis 10 I: Users estimate that a credibility-weighted tool would be more useful than an unweighted tool in an e-market place.

Hypothesis 10 II: Users estimate that a credibility-weighted tool would be more useful than an unweighted tool in an intra-company rating tool.

5.3.3 Method

5.3.3.1 Participants

The participants of Experiment II consisted of fifteen construction professionals working for three Bay Area general contractors. All of the participants were actively involved in bidding but they occupied four different positions (Estimators= 5, Project Managers=5 and Operations Managers=2, Project Engineers = 3). The participants were randomly assigned to the order in which the different rating tools were presented.

5.3.3.2 Procedure

In order to make the experiment as realistic as possible, the users evaluated a set of real bids from the subcontractors that had been hired to construct a San Francisco office building in 2001. For some of the trades, where the original subcontractor was less well known, I added an additional Bay Area subcontractor to increase the probability that the participants would know at least one of the subcontractors bidding for the trade. In total, the experiment involved twenty-six subcontractors bidding to perform the sixteen different trades that were subcontracted on the $3M office building.

The first step involved gathering a set of ratings of the subcontractors' performance. Prior to the experiment the participants received a survey asking them to rate the twenty-six subcontractors on seven different criteria using a ten-item Likert scale. The criteria were based on interviews with seven industry practitioners, as well as comments received during Experiment I. All the criteria involved an element of subjectivity to justify their weighting by rater credibility. Eleven of the participants returned this survey before the experiment. The remaining four rated the subcontractors after having completed the experiment. In the experiment, I had changed the names of the subcontractors in order to avoid recognition and to eliminate any order effects.

Once a set of ratings had been gathered, I contacted each participant individually and set up an appointment to perform the experiment. The experiment began with a ten-minute presentation that introduced the participant to the concept of AEC rating systems. Next the participant was asked to follow the instruction provided by the Microsoft Excel application that ran the experiment. The experiment was a within subject design with the type of tool as the within subject factor. It was carried out on an individual basis, using a personal computer at the participant's workplace, under the supervision of a member of the research team. Upon starting the application, the participant read an introduction saying that he or she was going to act as a project manager at a general contractor. The task was to compose a bid for the construction of a small office building and two different Internet-based rating systems were available to support his or her decisions.

Next, the participant saw a list of raters (corresponding to all the other participants that had rated the subcontractors) and was asked to identify whom of these raters he or knew, before using the McCroskey source credibility scale to rate a set of people. The people rated included the raters that the participant knew, one unknown project manager working for each of the three contractors participating in the study, and finally an unknown project manager working for an unknown contractor. Next the participants calibrated the rating tool by making pair-wise comparisons of the credibility of the different raters as described in Section 4.2.4.1. The participants evaluated divergent ratings from all possible (twenty-one) pairs of raters before the system performed an exponential regression. The user interface for this exercise is shown in Figure 7 in the description of TrustBuilder II.

After finishing the calibration, the participants evaluated the first half of the twenty-six low bidding subcontractors using either the credibility weighted or the unweighted tool before using the other tool to evaluate the remaining thirteen subcontractors. In the credibility weighted rating tool (corresponding to TrustBuilder II see Figure 8), the participant could see the overall ratings of the low bidding subcontractor, the identity of the raters, and the rater agreement. The unweighted rating tool, a simplified version of the credibility-weighted rating tool, showed the average ratings, and the agreement, along with the number of raters. In both tools the subcontractors' low bids were roughly equal to those of the original project. The two tools also displayed the bids of four competing subcontractors. A random function dynamically generated the competing bids, which were, on average, 15% higher than the low bid the user was evaluating.

It is also important to note that the name of the low bidding subcontractor had been changed. This prevented the participants from recognizing the subcontractors and thus evaluating them based on previous experience. In both tools the participants evaluated overall subcontractor quality, stated how confident they were in their judgment, as well as how comfortable they were hiring the subcontractor, and adjusted the subcontractor's low bid by adding a line item contingency. In the next step, those participants who had not yet rated any subcontractors were asked to rate the performance of the twenty-six subcontractors. They now had access to the real names of the subcontractors. Finally, the participants ended the experiment by filling out a questionnaire about their attitude towards rating systems.

5.3.3.3 Measures

Rater credibility was measured with the McCroskey twelve-item credibility scale and calculated following the procedure described above.

Relative Rater Weight was measured on a ten-item Likert scale using the User interface shown in described in Figure 7.

Model Fit was measured using the sum of squared errors in the pair-wise comparisons for the two models.

Factor Coefficients in Credibility Model: The coefficients for the four factors in TrustBuilder II (Know Rater, Same Organization, Trustworthiness, and Expertise) were estimated using Equation 9.

Rater Consistency: Similar to the Group Lens collaborative filtering model [17]. Experiment II used the Pearson coefficient of correlation to model rater consistency. The correlation of user i's and rater j's ratings are calculated as:

$$COij = \frac{\sum_{k,l}(R_{jkl} - \overline{R}_{il})(R_{ikl} - \overline{R}_{jl})}{\sqrt{\sum_{k,l}(R_{ilk} - \overline{R}_{il})^2 \sum_{k,l}(R_{ikl} - \overline{R}_{il})^2}} \qquad (12)$$

where R_{ikl} is user i's rating of subcontractor k on criteria l.

Bid contingency was measured with a single item, which consisted of a percentage number between -30% and 100%[9]. Bid contingency could reflect both the participants' assessment of the risk buffer that should be added to the bid, as well as the extra cost of managing an under-performing subcontractor.

Users' assessments of the qualification of the subcontractors was measured with a single question: "How qualified is X to perform this job?"

Users' confidence in their assessments was measured with a single item question ("How confident are you in your judgment?") referring to the user's estimation of subcontractor qualification.

Usefulness of the model: The experiment measured the user's opinions of the usefulness with four single item questions:

"How useful would a rating tool that weights all ratings the same be in an e-market place?"

"How useful would a rating tool that weights ratings by rater credibility be in an e-market place?"

"How useful would a rating tool that weights all ratings the same be in an intra-company supplier evaluation system?"

"How useful would a rating tool that weights ratings by rater credibility be in an intra-company supplier evaluation system?"

5.3.4 Results

5.3.4.1 Significance of factors in credibility model

I performed a bootstrap analysis to analyze the significance of the four factors (Know Rater, Same Organization, Trustworthiness, and Expertise) proposed in the TrustBuilder II version of the credibility-weighted model. The bootstrap is a computational method for obtaining an approximate sampling distribution of a statistic. Since, it is conditional on the observed data, the bootstrap enables us to

[9] An interviewee during the pre-study suggested the possibility of allowing the users to enter a negative risk buffer or contingency. In the end, no participant took advantage of this possibility during the experiment.

estimate confidence intervals, even though we do not know a parameter's exact distribution. To enable the bootstrap analysis of the credibility model, I created a small S-PLUS [169] program that randomly samples (with replacement the) a set of fifteen users. The program then performs the exponential regression (to estimate the coefficients in Equation 9) based on the 315 comparisons provided by the fifteen users in the sample. The program performed this procedure 2000 times to obtain statistically significant estimates. Figure 15 shows that all four factors were positive within a 95% confidence interval in the bootstrap analysis. It is important to point out that the different factors are by nature correlated. The fact that the user knows a rater increases the likelihood that the two will work for the same as organization and makes it more probable that the user will find the rater trustworthy and competent. Still, the result, showing all the factors to be positive within a 95% confidence interval, provides evidence that all of the four factors in the TrustBuilder II model contribute to the estimation of rater credibility (Hypothesis 7). More specifically, the results indicate that the two classical factors in source credibility theory (e.g., [24]), perceived expertise and trustworthiness, contribute to the prediction of rater weights in an AEC rating application.

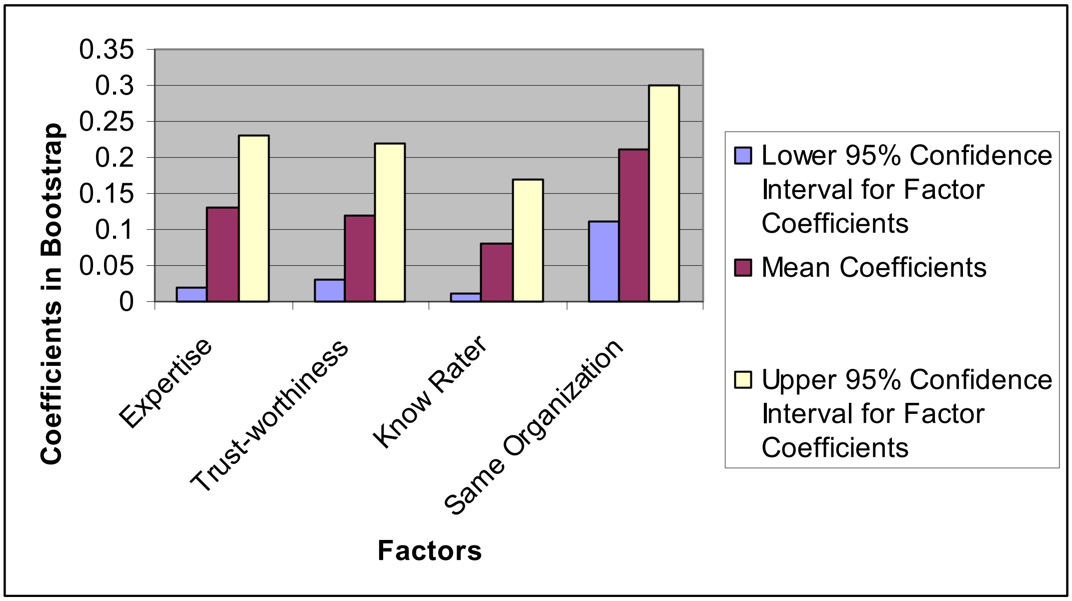


Figure 15 Results from bootstrap analysis which show coefficients of factors in exponential regression of rater weights. As shown all coefficients are positive in the 95% confidence interval.

The results imply that all factors in the model (including perceived expertise and trustworthiness from source credibility theory) are significant predictors of rater weight.

The four factors had been normalized to z-scores and therefore had equal variance. The size of the coefficients in the bootstrap therefore provides some indication of the relative importance of the different factors. As shown in Figure 15, the results suggest that the factor, which has the highest influence on rater weight, is whether the rater and user work for the same organization. However, it is important to note that the large spread in the coefficients prevents any statistical verification of this finding.

5.3.4.2 Ability to predict rater weights

To evaluate the two models' ability to predict rater weights, I performed two different tests. As a first step, I performed a maximum likelihood ratio test using the same methodology as in Experiment I. Similar to Experiment I, the maximum likelihood ratio test compared the errors of an unweighted model to the errors of a credibility-weighted model that had been calibrated at the individual user level. As Figure 16 shows, the average squared error in the credibility-weighted model (0.017) to be considerably smaller than in the unweighted model (0.071). The maximum likelihood ratio test, which takes account of the different degrees of freedom (60 vs. 0) of the two models, shows that this difference is significant (p< 0.0001).

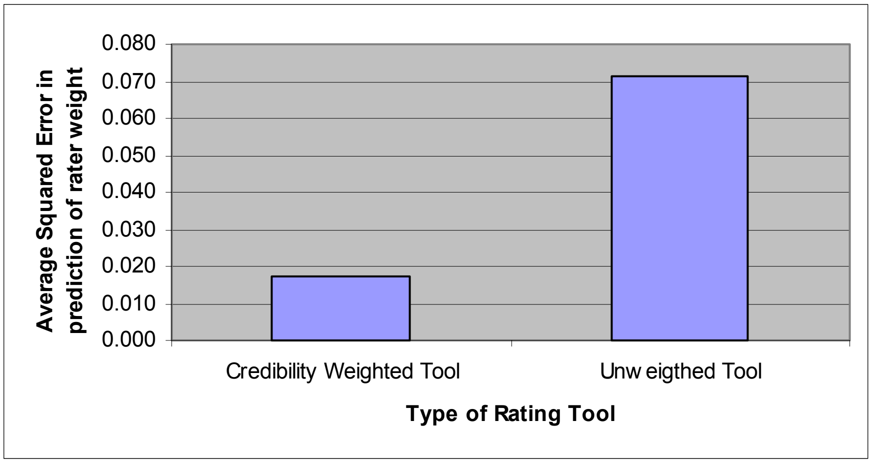


Figure 16 When predicting the users' assignments of rater weights the average errors squared errors were considerable smaller in the credibility-weighted model (0.017) than in the unweighted model (0.071).

To confirm this result, I created another small SPLUS program which performed a cross validation [169]. To obtain more stable results in the cross validation algorithm, I estimated the coefficients across all users. The cross validation not only compared the performance of the credibility-weighted model to that of an unweighted model but also to the performance of a third model. This third model was similar to the credibility-weighted model but contained only one binary variable, which modeled whether the user knew the rater. (I will refer to this model as "Know Only.") In the Know Only model, the credibility of rater j from user i's perspective (C_{ij}) was thus modeled as:

$$C_{ij} = \exp(\alpha + \beta * KR_{ij}) \qquad (13)$$

Where
α: constant set to -1
β: coefficient in exponential regression
KR: z-score normalized for each user modeling whether the user i knows the rater j.

The cross validation was run in a small S-plus program that first created a training sample by selecting (without replacement) twelve of the fifteen users. The algorithm then fitted the credibility-weighted and Know Only models based on the 252 weight estimations provided by these twelve users in the training sample. In the next step, the models predicted the weights on a test sample consisting of the 63 data points provided by the remaining three users. Finally, the program calculated the differences between the means of the squared errors generated from the three models (the two trained models model and the unweighted model) when they are applied on the test sample. This procedure was repeated 2000 times and the results are displayed in Figure 17, which shows the mean, along with the 90%, and 95% confidence intervals for the differences in the means of the squared errors on the test sample. The unweighted model clearly has a higher mean error than the credibility weighted model. In fact, in all of the 2000 simulated cases, the mean error was smaller for the credibility-weighted model than for the unweighted model (i.e., the difference in mean squared errors between the two models was positive.) This result is consistent with the outcome of the maximum likelihood ratio tests for both Experiment I and Experiment II, which both showed the credibility-weighted model to be better than an unweighted model at predicting rater weights.

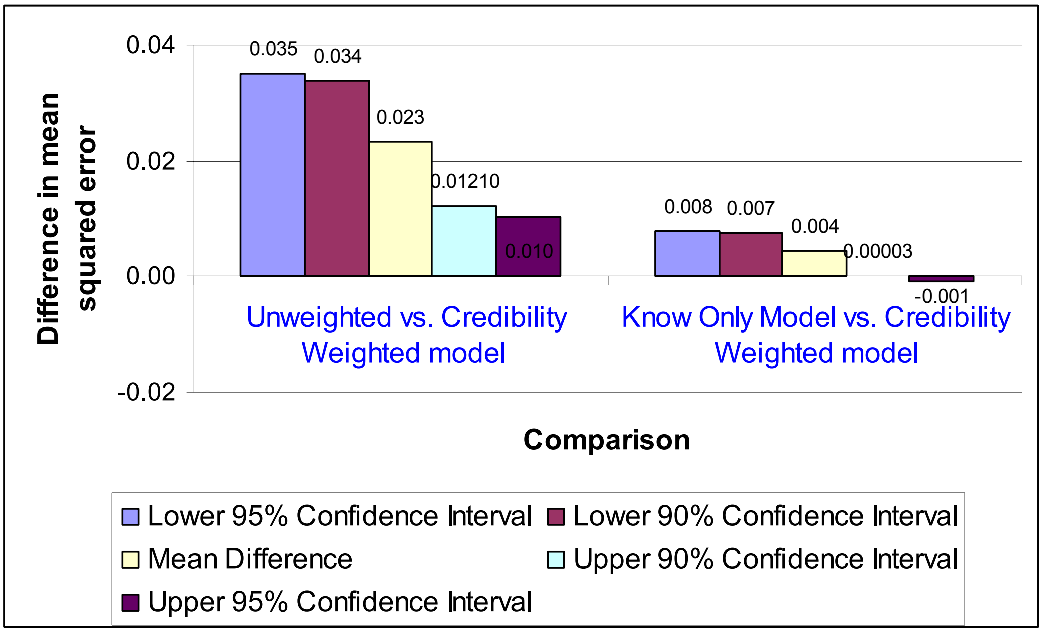


Figure 17 Differences in mean of squared errors on test sample in cross-validation. The differences in error between the unweighted and credibility weighted models are positive in all 2000 simulated cases, which indicates that the credibility model is better at predicting rater weight. The differences between the Know Only and Credibility-weighted models are smaller but positive within a 90% confidence interval.

The cross validation shows that the credibility weighted model also fares better than the unweighted model on independent data. As a result, I conclude that this research project has provides strong evidence that a credibility-weighted tool is better than an unweighted (constant) model at predicting the relative weights users attribute to different raters.

The difference in terms of mean squared errors between the Know Only model and credibility-weighted model is much smaller. All values within the 90% confidence interval are positive whereas the 95% confidence interval contains some negative values. This result indicates that it is not enough only to know whether a user knows the rater. Instead there is a need for a larger model, which includes credibility measures as well as information about rater's organizational affiliation.

5.3.4.3 Prediction of Consistency

I tested two regression models to study the extent to which the factors in the credibility-weighted model predict rater consistency. The first model used the four factors of TrustBuilder II credibility-weighted model (“Rater Expertise”, “Rater Trustworthiness”, “Know Rater” and “Same Organization”) as predictors of rater consistency. The resulting regression model (see Figure 18) was significant ($p<0.001$) but also involved a lot of unexplained variance (Adjusted R-square = 0.13). Moreover, the regression model indicated “Know Rater” to be the only significant factor. I therefore also tested a reduced model, which used only that factor to predict consistency. The result was a marginal decrease in R-square (from 0.130 to 0.127) but, given that the smaller model only contains one factor instead of four, it is clearly a better model.

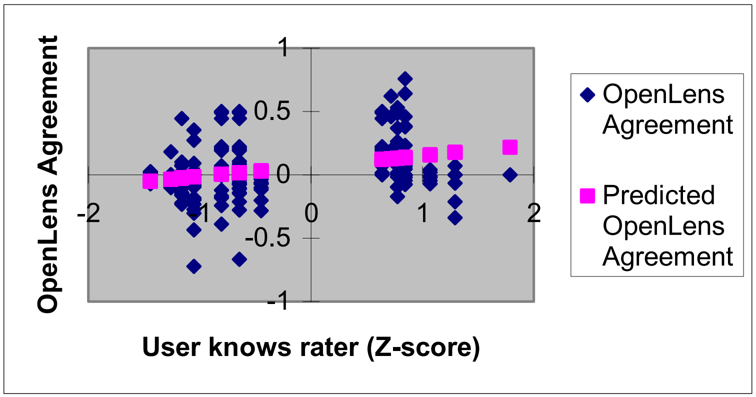


Figure 18 Linear Regression of Rater Consistency calculated using Group Lens measure and z-score of whether the user knows the rater

The results indicate that for predicting rater consistency the only factor of importance is whether the rater knows the user. Neither the user’s perceptions of rater trustworthiness and expertise, nor whether the rater and the user work for the same company, seem to be important predictors. This result is not very surprising. For a user’s subjective judgment of a rater’s trustworthiness and expertise to be correlated with the consistency between the rater’s and the user’s ratings, one would expect that the user had seen the rater’s ratings prior to making the credibility judgments. If the estimator Chuck Numbers notices a big discrepancy between his and fellow estimator Jim’s ratings, he is likely to downgrade his judgment of Jim’s expertise and/or trustworthiness somewhat. Although expected, the lack of correlation between a user’s estimate of rater credibility and the consistency between the user’s and rater’s ratings is important. The finding indicates that the measures of trust calculated through pure data analytical algorithms (e.g., [17-19, 65, 121]) are less likely to make intuitive sense to the users. These rating models have been designed to function primarily in large online communities where the vast majority of the users are anonymous (eBay, for example). In Business-to-Business e-commerce communities where the market participants comprise, for example, construction industry estimators and project managers, the likelihood that a user knows a rater will be much higher. In addition, in a rating system that incorporates internal as well as external ratings, the likelihood that a user will know a rater increases even more. Chuck Numbers would probably be surprised if a collaborative reputation mechanism gave his trusted friend and coworker Jim a low weight.

This result, that people who know each other have more consistent ratings, supports an opinion expressed by several of the participants of the experiment who stated that “the construction industry is a `people business’”. In other words, the relationship between the general contractor and the subcontractor is contingent on the individual contact persons at the two organizations. It is fair to assume that two users who know each other are likely to work with the same contact person at the subcontractor. Another explanation is that people who know each other discuss the subcontractors and therefore share the same view of the subcontractors’ reputation. If B tells C that PaintA did a bad job, this will probably affect C’s opinion of PaintA. This effect is similar to Kilduff and Krackhardt’s observation that “the performance reputation of people with prominent friends will tend to benefit from the public perception that they are linked to those friends.” It is likely that this “basking in reflected glory effect”, could manifest itself also in the context of AEC bidding. For instance, an estimator who perceives that a highly regarded project manager frequently hires a subcontractor could tend to regard the subcontractor’s qualities in more favorable light.

The evaluation also shows that there is a considerably amount of variance that is not explained by the model. This result can be explained in several ways. Firstly, there may not be enough data to train the Group lens consistency measures properly.

In total, the fifteen users provided 815 ratings distributed over twenty-three of the twenty-six subcontractors. Furthermore, subcontractor performance is subject to high variability. Factors such as weather, manpower and management by the general contractor all influence subcontractor performance. As one project manager expressed when rating the performance of a small electrical subcontractor, and referring to the subcontractor's performance on a recent job: "Acme Metalworks they struggled. They were good but they did not have enough manpower to put on the job." This subcontractor received better ratings from other participants who presumably had hired it for smaller jobs.

5.3.4.4 Variance of Bid evaluations

To test whether users varied their overall evaluations more using the credibility weighted tool than with the unweighted tool (Hypothesis 9I), I again applied the Wilcoxon matched-pairs signed-ranks test, deeming the form of the yielded data suitable following the criteria and procedures set out in Cohen and Holliday [167]. Figure 19 shows the variance of the users' overall evaluations of subcontractor performance with each of the two tools. The Wilcoxon matched-pairs signed-ranks test (W+ = 94.50, W- = 25.50, N = 15, p <= 0.04791) showed that users varied the contingency more using the credibility-weighted tool (Hypothesis 9 I).

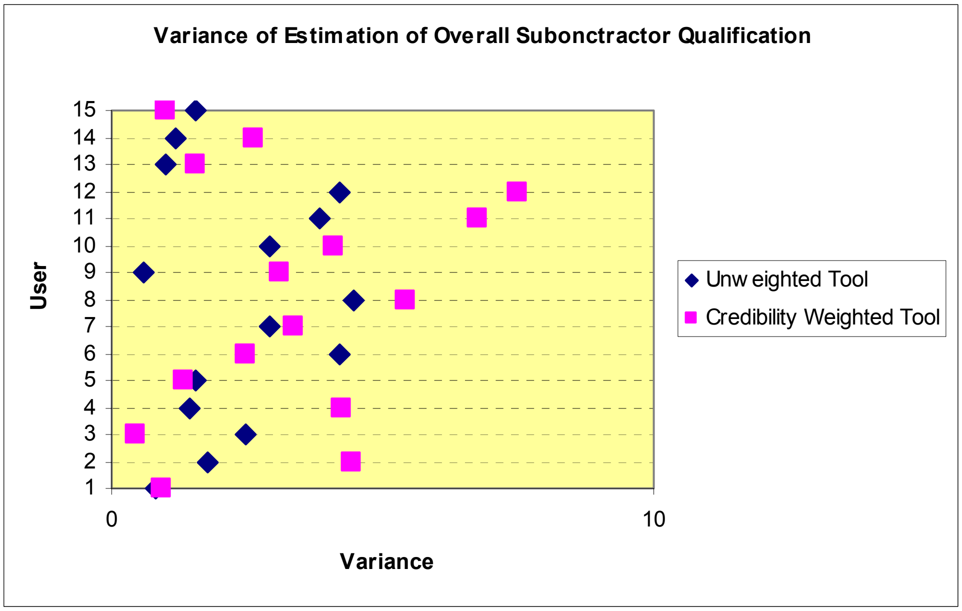


Figure 19: Variance of each user's evaluations of overall subcontractor qualification using the unweighted and credibility weighted tools. The variance is higher for the credibility-weighted tool than for the unweighted tool (p < 0.05).

The results show that AEC practitioners do vary their evaluations of subcontractor performance more when using the credibility weighted rating tool than when using the unweighted tool. This outcome suggests that the users trust the data supplied by the credibility weighted tool more than the information supplied by the unweighted tool. This is especially the case when it is credible peer raters who are supplying the ratings as this will lead to increased user trust. Subjective data collected during the experiment support this hypothesis. To cite a participant, who was using the credibility-weighted tool: "*The overall rating of Trojan Electric is 9 out of 10 and Chief Estimator Maldini is among the raters. I go with the 9*." (Several other participants spontaneously made similar statements.) There exists a similar explanation to the behavior of two of the users (3 and 6 in Figure 19) who varied their evaluations more when using the unweighted tool. It turned out that users 3 and 6 only knew one of the raters of the subcontractors.(All other users knew at least three of the raters.) In addition, the known rater had only rated about half of the subcontractors that the two users evaluated with the credibility-weighted tool. Using the credibility-weighted tool, the two users could see that the overall ratings were based on information from a set of unknown raters with low perceived credibility.

Consequently, they felt less inclined to let the overall ratings influence their evaluations. Since the unweighted tool, did not show the raters' identity, it was not apparent to the two users that the overall ratings were based on information from raters with low credibility. It is therefore logical that they varied their evaluations more using the unweighted tool. On the other hand, I could not find an explanation to why user 15 varied his evaluations more using the unweighted tool. After completing the experiment he said that: "I weigh people I know higher than people I don't know. " It is not apparent why this attitude was not reflected in his behavior when evaluating overall subcontractor qualification.

To conclude, the results provide evidence that credibility information adds value when AEC practitioners evaluate subcontractors. The results are also consistent with those of Experiment I, which showed that participants varied bid contingency more when using the credibility-weighted tool.

5.3.4.5 Bid Contingency

In Experiment I, the users varied their bid contingency decisions more when using a credibility weighted-tool (Hypothesis 2). Experiment II investigated whether it was possible to replicate this result in an environment where more factors than the subcontractor ratings were allowed to vary. This investigation followed the same procedures as in Experiment I, but the results were non-conclusive. There were no significant differences in either direction. Seven of the users varied the contingency more when using the unweighted tool while six varied more when using the credibility-weighted tool. Two of the users, both of whom were estimators working for a general contractor, who mostly does competitive bidding, did not add any contingency at all. One of them explained that: "*I work in a competitive environment. Adding a risk buffer [contingency] may be a very good idea. But if I do it my competitor [who does not add any contingency] will bid lower and get the job.*" Several users expressed similar attitudes and thus only added contingency when the subcontractor's bid and ratings deviated significantly. Interesting to note was that estimators, on average, added less contingency (Mean = 2.94) than did project managers (Mean = 6.34). One explanation is that, on a competitive job, the estimator's primary goal is to bid lower than her competitors in order to win the contract. Of course she does not want to win the bid at a loss, but she will still not add any extra buffer to a low bid sub (or choose a subcontractor other than the low bidder) unless there is a compelling reason to do so. If the company wins the contract, the job is handed over to a project manager who is responsible for managing the hired subcontractors. A project manager, on the other hand, would not mind adding a risk buffer to a bid since this will only increase the probability that the project will stay within budget.

In order to further investigate the rationale behind the bid decisions, I performed a linear regression of bid contingency. The full regression model tested the following independent variables as predictors of contingency:

- *Bid Amount*: What is the absolute size of the bid? Users can be expected to adjust the bid more if the dollar value of the bid is large. Given the large spread in bid sizes, which range between $5,000 and roughly $400,000, I decided to use the logarithm of bid size,
- *Distance from second lowest bid*: How much lower is the evaluated subcontractor's low bid compared to the second lowest bid?
- *Type of rating tool* is a binary measure, which is coded 0 for the credibility-weighted tool and 1 for the unweighted tool.
- *Overall Qualification*: The overall qualification of the bidding subcontractor estimated by the user corresponds to the user's estimate of the overall ratings of the subcontractor.
- *Agreement* is the rater agreement, as shown in the user interface.
- *Number of Raters* who had rated the subcontractor.

Prior to performing the regression, I normalized the contingency measures, as well as the independent variables, to z-scores. The normalization of the contingence measures was performed at the individual user level in order to compensate for individual differences due to, for instance, the user's type of job. The normalization of the independent variables, on the other hand, took place over the entire data set, which facilitated the comparison of the variables' relative importance, giving each factor equal mean and variance. For two reasons, the final regression model included data from only eight of the participants. Firstly, I classified the two estimators who did not vary their contingency at all as outliers and their data was therefore not used in the regression. More unfortunate, a programming error in the Excel application that ran the experiment erased the competing bids for five of the users, for whom it was thus not possible to calculate the variable “Distance from Second lowest bid.” However, the transcripts of user behavior do not show any significant differences for these five users compared to the rest of the participants.

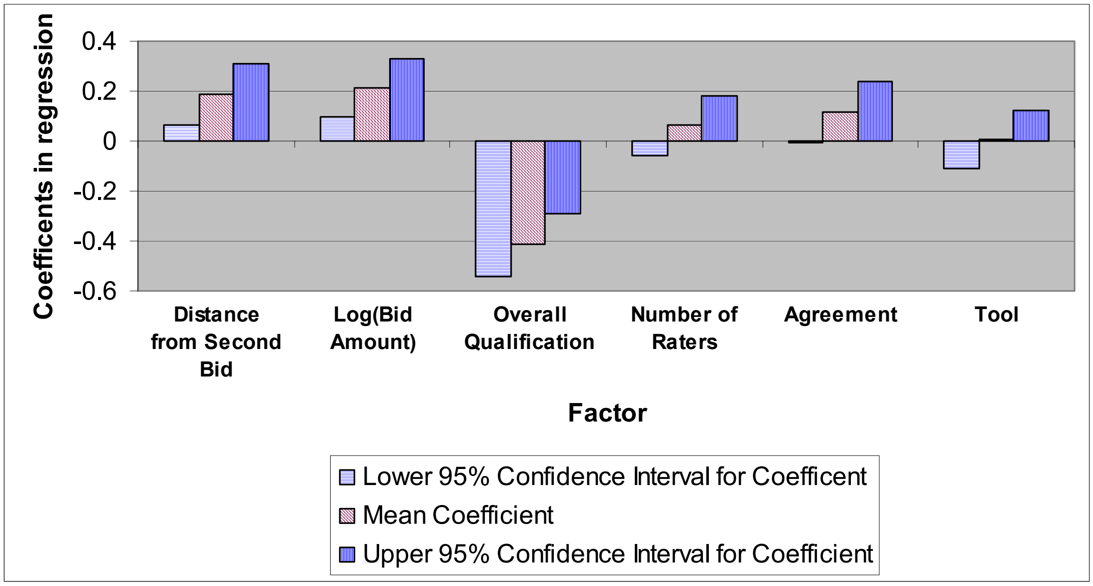


Figure 20 Coefficients in regression of bid contingency. Overall qualification, Bid Amount, and Distance from second bid are all significant (95% confidence)

The linear regression of bid contingency (Figure 20) showed only three significant factors at a 95% confidence level. The first significant factor was the overall qualification of the subcontractor as estimated by the user, which shows that the two measures “bid contingency” and “overall qualification” are correlated. However, since the adjusted R-square of the model was only 0.19, I conclude that the overall rating only partly explains bid contingency.

Bid Amount also turned out to be statistically significant. The larger the dollar value of the subcontract is, the riskier the contract becomes, which increases the user’s inclination to add a risk buffer. For small contracts, participants did not feel that it was necessary to add any contingency since they considered the financial exposure to be minimal. One participant explained this rationale as follows: "*HVAC units only $10 K. I don't have to add any contingency here*".

The regression model generated the distance from second lowest bid as a significant factor, which confirms statements from several participants who explained their bid strategy in, for example, the following way:

"*I look at the bid and compare it to the other bids. If his [the subcontractor's] numbers are off I bump up the bid [add contingency] so that it is in level with the second lowest bidder*."

As mentioned above, the R-square measure of fit for the regression model was relatively low (0.19). I attribute this result to the fact that several important factors, which could not be represented by ordinal variables, were missing in the regression model. One such factor is the type of trade. Decision makers perceive complex trades, Controls, for instance, to involve more risk than simpler ones trades such as Painting. "*Nobody likes controls guys*" "*Painting, I don't care. What can go wrong?"* The bid amount factor partly served to control for the impact of the type of trade. Complex and risky trades (e.g., HVAC) are in general more expensive than simpler, more commoditized trades such as carpet installation. The regression model also excluded the participant's individual strategy for adding contingency, which added considerable variance to the results. Converting the contingency measures to individual z-scores only partly models the individual user's strategy.

It is interesting to note that the type of rating tool does not seem to influence the absolute level of bid contingency added to a bid. However, this outcome does not contradict Hypothesis 4 II, which states that users would *vary* bid contingency more with the credibility-weighted tool. Instead, the regression model showed Bid size and Distance from second bidder to be two determinants of bid contingency. Since these two factors are not related to the type of rating tool, I further investigated Hypothesis 4 II by constructing a second regression model, which controlled for the variance added by the two factors. As a first step, I regressed bid contingency (normalized for each user[10]) against log (Bid Size) and Difference from second Bid. The errors in this regression constitute an estimate of the bid contingency, which is not attributed to

[10] Bid Contingency was normalized for each user by dividing by the mean contingency that the user applied to the bids.

Bid Size or Difference from second bidder. I then calculated the variance of this measure when the participants were using each of the two tools. The result showed that, on average, the participants had a slightly higher variance (0.92) when using the credibility-weighted tool than they did when using the unweighted tool (0.85), but that this difference is to small to be statistically different.

The results above indicate that, given the conditions of Experiment II, the type of rating tool does not significantly affect user behavior regarding the task of adding bid contingency to subcontractor bids. In my opinion, the reason for this result is that bid contingency is a highly individual process, which, as shown in the regression model, is a function of several factors. When a user is simultaneously considering factors such as the size of the bid, the distance from the other bids, the ratings of subcontractor on six different rated criteria, the rater agreement, and the type of trade, participants pay less attention to the type of rating tool. It is also important to point out that in their current practice none of the participants uses a rating tool to evaluate bids. Therefore, a reasonable interpretation is that, when adding contingency, the users primarily considered the criteria they were familiar with, before considering the type of tool they were using.

5.3.4.6 Confidence in evaluations

To check whether the users were more confident using the credibility tool (Hypothesis 3), I performed a linear regression of confidence in ratings, using the factors listed below as independent variables. In order to use data from as many participants as possible, the model did not include Distance from Second Bid as a factor, since data for this factor was unavailable for five of the participants. However, a regression that included also this factor, and was then limited to the data provided by eight users, generated basically the same results. The regression model included the following predictors of user confidence:

- Log (Size of Bid)
- Type of Rating tool
- Overall Qualification
- Agreement

- Number of Raters

I classified two of the fifteen users as outliers (whose data were not used in the regression), since they entered constant confidence for all twenty-five bids. As result, I could normalize the factors in the regression model to z-scores. The normalization partly accounts for individual user strategy and behavior when assessing overall qualification and confidence.

The regression model generated three significant predictors of user confidence. Figure 21 shows that the type of rating tool is a significant predictor of rater confidence. The negative coefficient (0.10, t-test: N=325, $p<0.05$) indicates that users are more confident when using the credibility-weighted tool (coded as 1) than when using the unweighted tool (coded 0). This result is consistent with the results generated in Experiment I, and shows that the use of a credibility-weighted rating tool increases the user's confidence in the accuracy of the information.

As Figure 21 shows, another factor that contribute to user confidence is Overall qualification. The more qualified the user estimates the subcontractor to be, the more confident she becomes in her judgment. One explanation for this phenomenon is that while a high overall rating (e.g., 9.4) from a set of raters indicates that the subcontractor has consistently performed well, more than one factor might result in an average rating. An overall rating 6.2, for instance, could be composed of a set of ratings close to 6, but is more likely to be the result of a mix of high and low ratings. The purpose of the agreement index was to help users differentiate between consistent and non-consistent ratings, but, as shown in the regression, it does not seem to have influenced user confidence.

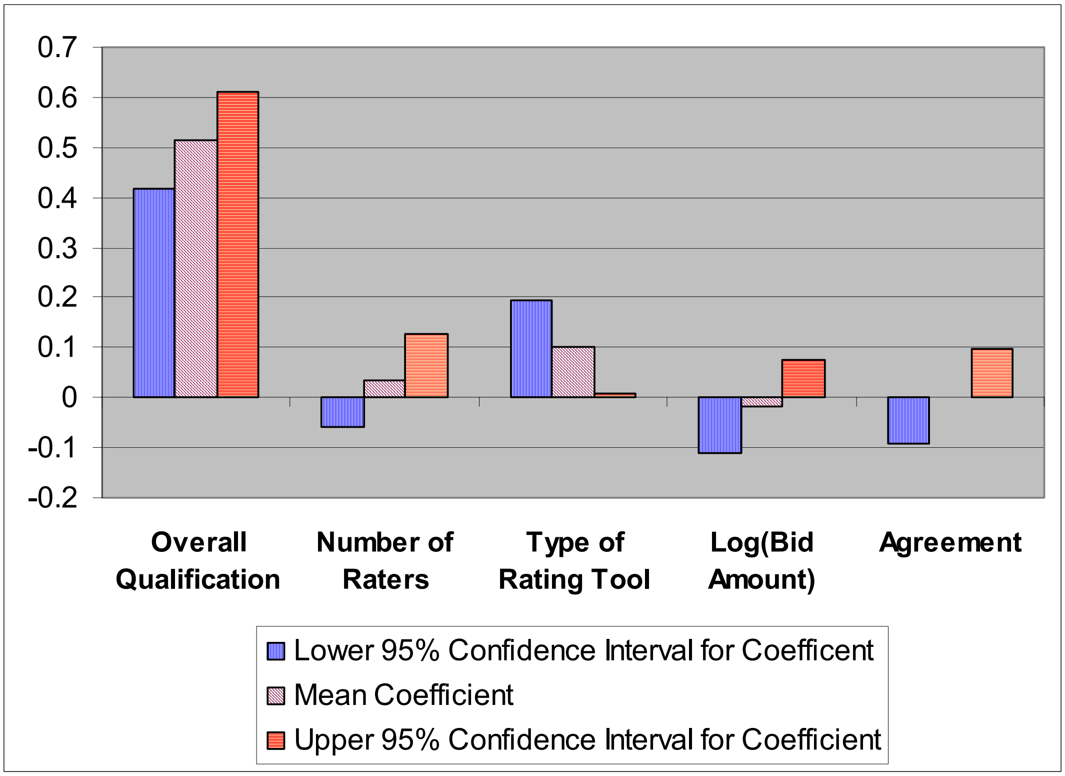


Figure 21 Coefficients in regression of user confidence. The positive coefficient (0.10, t-test: N=375, p<0.05) indicates that users are more confident when using the credibility-weighted tool (coded as 1) than when using the unweighted tool (coded 0).

We should note that, since each user performed twenty-five evaluations, there the observations are not independent. To deal with this problem, I constructed a second model where the user was the unit of analysis instead of the rating. However, this approach also decreases the number of observations from 325 to 13, limiting the chances of obtaining statistically significant results. In this case, the fact that overall qualification is the main determinant of rater confidence further complicates the analysis. I therefore constructed a regression mode,l which controlled for the impact of overall qualification. As a first step, I regressed confidence against overall qualification. The errors in this regression constitute an estimate of the confidence, which is not attributed to overall qualification. I then calculated the average of this measure (user confidence given overall subcontractor qualification) when the participants were using each of the two tools. The result showed that, given overall subcontractor qualification, the participants were, on average, more confident (0.10) when using the credibility-weighted tool than they were using the unweighted tool (-0.10, t(12) =1.52, p<0.1). This result confirms the conclusion of the regression analysis (that the user's confidence in the accuracy of the information is higher in the credibility weighted than in the unweighted tool) holds also when user is the unit of analysis.

5.3.4.7 Usefulness of the tool

Finally, Experiment II measured the extent to which the participants found the two rating tools to be useful. I performed a pair-wise t-test of the survey answers to compare the user's ratings of tool usefulness in an e-market place and an intra-company rating tool respectively. It is standard practice to apply this test to analyze differences in the means of variables, which are measured using standard Likert scales [168]. The underlying assumption behind the paired t-test is that the variables are normally distributed. For the e-market place, users found the credibility-weighted tool more useful (M=8.73, SD=1.03) than the unweighted tool (M=7.53, SD=2.45, N=15, paired t-test: p<0.05.) Also for the intra company settings, the participants estimated the usefulness to be higher for the credibility-weighted tool (M=8.87, SD=1.13) than for the unweighted tool (M=7.67, SD=2.19, N=15, paired t-test: p<0.05.) To confirm these results I also performed a bootstrap analysis, which does not rely on assumptions about the shape of the distributions. The object of the analysis was the difference between the participants' estimates of the usefulness of the credibility-weighted and the usefulness of the unweighted tool. A positive difference indicates that the user finds the credibility-weighted tool to be more useful. For example, if Jane Estimator rates the usefulness of the credibility-weighed tool to be 9 and the usefulness of the unweighted tool to be 7, the difference will equal 2. The bootstrap analysis randomly selected, with replacement, the differences for fifteen of the users and calculated the mean of these 15 differences. Based on 1000 samples, the bootstrap analyses were consistent with the paired t-test. As Figures 20-21 show, the 95% confidence intervals for the mean of the differences were strictly positive for the e-market place ([0.20, 2.33]), as well as the intra company setting ([0.20,2.27]).

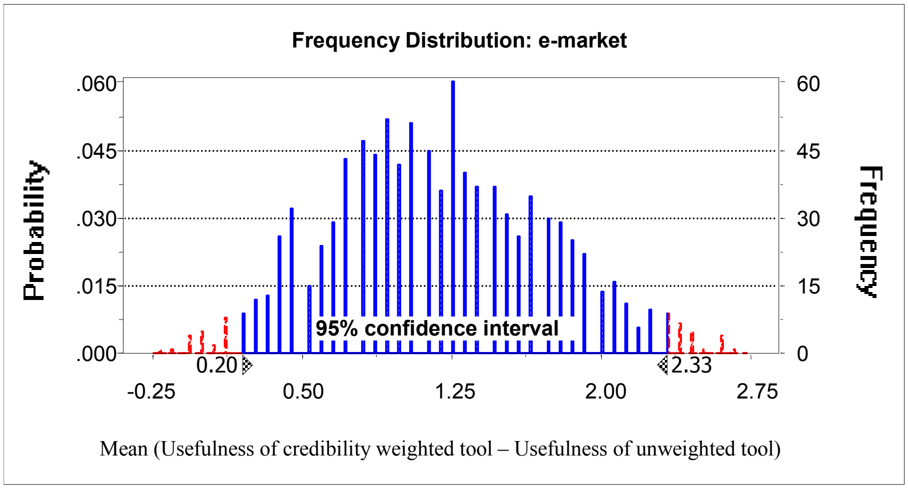


Figure 22 A bootstrap analysis investigated the participants' estimates of the usefulness of the two tools in an e-market place. The object of the analysis was the difference between the participants' estimates of the usefulness of the credibility-weighted and the usefulness of the unweighted tool. The 95% confidence interval for the mean of this difference was strictly positive ([0.20, 2.33]), which shows that the users estimated the credibility-weighted tool to be more useful in an e-market place.

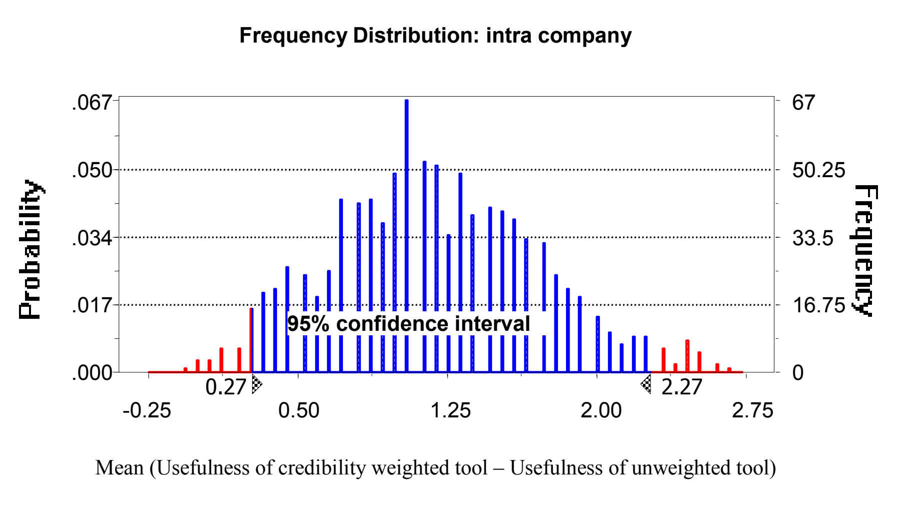


Figure 23: A bootstrap analysis investigated the participants' estimates of the usefulness of the two tools in an intra company setting. The object of the analysis was the difference between the participants' estimates of the usefulness of the credibility-weighted and the usefulness of the unweighted tool. The 95% confidence interval for the mean of this difference was strictly positive ([0.20, 2.33]), which shows that the users estimated the credibility-weighted tool to be more useful in intra-company setting.

I therefore conclude that, on average, the industry practitioners found the credibility-weighed tool to be more useful than the unweighted tool.

These results illustrate the potential use of credibility weighted rating tools in two different settings. The first setting is an e-market place where knowledge is exchanged across organizations. As we would expect, users appreciate the opportunity of giving different weights to users within their own and other organizations. A credibility-weighted rating system could also be deployed internally, within the organization of a large contractor. The results show that the participants also found it useful to differentiate between different types of users within their own organization. An estimator can be expected to find a close friend with extensive industry experience to be more credible than a newly hired project engineer whom he has never met, even though the three work for the same company.

5.4 Summary and discussion of the results from Experiment 1-2

In this section, I summarize the results of Experiment I and 2, and discuss to what extent they answer the fundamental research question:

How can source credibility theory support rating systems in the procurement of AEC services?

The following paragraphs discuss the research question in terms of its two components: operationalization and added value.

5.4.1 Operationalization

Both experiments investigated the first part of the research question:

How is it possible to operationalize source credibility to support the calculation of weights that are based on rater identity in an AEC rating tool?

This research project shows convincing evidence that the proposed methodology in TrustBuilder I and II to operationalize source credibility theory can support rating tools for AEC bidding. Table 19 shows a summary display of the research hypotheses, which refer to operationalization along with the outcomes of Experiment I and II.

Table 19: Summary of the results concerning operationalization from the two experiments. Two different versions of the TrustBuilder methodology to operationalize source credibility theory fared better at predicting rater weights than an unweighted model.

Hypothesis	**Evidence in Experiment I**	**Evidence in Experiment 2**
Rater Weights I *A credibility weighted model will better model the rater weights (calculated at the individual user level) expressed by users in pair-wise comparisons than an unweighted model.*	Maximum likelihood Ratio Test (p<.0001)	Maximum likelihood Ratio Test (p< .0001)
Rater Weights II *A credibility-weighted model will better model the rater weights (calculated at the aggregate level) expressed by users in pair-wise comparisons than an unweighted model.*		Cross Validation (p< .001)
Factor Coefficients *The factors used in the credibility weighted model all influence rater weight.*		Bootstrap Analysis (p<..05)

The two experiments tested two different versions of the TrustBuilder methodology to operationalize source credibility theory, using data from two different sets of users. I argue that the consistent outcomes from the experiments provide evidence for the generality of the method.

In Experiment I, the credibility weighted model of TrustBuilder I fared better than the unweighted model in terms of predicting rater weight at the individual user level. I was able to repeat this result in the second experiment, which involved the more refined model of TrustBuilder II, as well as more experienced participants.

A more extensive validation of the results of Experiment II also generated similar results when the data had been aggregated for the entire set of industry practitioners. First of all, a cross validation consistently generated larger error for the unweighted tool on the test sample. This outcome shows that that the superior performance of the credibility-weighted model was not caused by over-fitting the model. Furthermore, a bootstrap analysis showed the coefficients in the TrustBuilder II model to be positive. This result indicates that the two fundamental factors of source credibility, perceived expertise and trustworthiness, are both significant predictors of weight in the credibility-weighted model.

The findings of the two experiments support the hypothesis that it is possible to operationalize source credibility in order to create a model for calculating rater weights that is more accurate than a standard unweighted model. Furthermore, the proposed TrustBuilder model provides an example of a methodology to operationalize source credibility theory that is superior to a standard unweighted model. Evidence that the proposed operationalization is valid is not only an important research finding in itself, but also a prerequisite for any discussion of the added value provided by a source credibility weighted model in the context of AEC e-bidding.

5.4.2 Added Value

Both experiments also investigated the second sub-question of this research project:

How can a rating system based on source credibility theory add value in the process of evaluating AEC subcontractors?

The results provide evidence that a credibility-weighted rating system can add value in the context of electronic bidding in construction. This conclusion is supported by the fact that the two experiments found significant differences between a credibility-weighted and an unweighted model in terms of three measures: variance of overall qualification, confidence, and usefulness. I argue that, by knowing that the ratings are filtered by rater credibility, users are more confident about their evaluations and hence allow the ratings provided by the tool to influence their decisions to a larger extent. I will now compare and discuss the results, which are summarized in Table 20, in more detail.

Table 20 Summary of results regarding added value of source credibility in the context of AEC rating. The results from the two experiments consistently show that a credibility-weighted tool adds more value to an AEC practitioner's bid evaluation process, than does an unweighted tool.

Hypothesis	Evidence in Experiment I	Evidence in Experiment II
Variance of subcontractor evaluations I *Users will vary their overall ratings of subcontractor qualification more when using the credibility weighted rating tool than when using the unweighted rating tool.*	Not Applicable	Wilcoxon matched-pairs signed-ranks test: p < .05)
Variance of Subcontractor Evaluations II *Users will vary the contingency added to bids more when using the credibility weighted and the unweighted rating tools.*	Wilcoxon matched-pairs signed-ranks test: p < 0.005	No significant differences found.
Confidence in Ratings *The use of a credibility-weighted relative to an unweighted tool results in increased user confidence in their judgments of overall performance.*	t-test of average confidence: p<.05	t-test in regression analysis: p<.05 t-test of average user confidence given overall qualification: p<.1
Usefulness (electronic market place) *Users estimate that a credibility weighted tool would be more useful than an unweighted tool in an e-market place*	Not Applicable	t-test: p<.05
Usefulness (intra company) *Users estimate that a credibility weighted tool would be more useful than an unweighted tool in an intra-company rating system*	Not Applicable	t-test: p<.05

5.4.2.1 Impact of a rating tool on evaluations of subcontractor performance

When comparing and analyzing the results from the two experiments we must keep in mind that in the first experiment the users 1) were given information limited to the subcontractor ratings to support the evaluations; 2) and made only one evaluation of subcontractor performance (bid contingency). In the second experiment, the participants had access to information other than the ratings (e.g., Distance from second bid, Size of bid), which supported their evaluations. To control for the influence of these factors, I measured overall qualification as well as bid contingency. We also have to take into account the relative inexperience of the users in Experiment I regarding the task of evaluating subcontractors. As a result of this inexperience, they were more likely to base their bid contingency decisions on the information provided by the two rating tools rather than on any well-established strategy. None of the users in Experiment I followed the above-described principle, for example, of never adding any contingency to subcontractor bids. I therefore argue that the Bid contingency measure of Experiment I, in practice turned out to be equivalent to the Overall qualification measure of Experiment II. Therefore, we would also expect the results for Bid contingency in Experiment I, and overall qualification in Experiment II, to be consistent across the two experiments. Both experiments provide evidence that users will vary their evaluations of subcontractor performance more using the credibility-weighted tool than with the unweighted tool. The results support the argument that the use of a credibility-weighted rating tool increases the users' trust in the information and hence also their willingness to let the ratings influence their evaluations. This finding is important because, ultimately, the added value of a rating system is that it helps decision makers differentiate between suppliers in terms of quality.

Still, it is important to point out, that the lack of impact on bid contingency in Experiment II indicates that when it comes to directly influencing general contractors' decisions in the context of AEC competitive bidding, the type of rating tool appears to be of secondary importance. As I have shown, the type of rating tool does affect the user's evaluation of the expected performance of the subcontractor, making her vary this decision more. However, the results of Experiment II show that the expected performance of the subcontractor (independent of which tool is used) only partly determines the user's bid contingency decision. The general contractor, evaluating the bid, may believe that the subcontractor is likely to perform poorly, but estimates that, in view of the competition, it cannot afford to add any risk buffer.

5.4.2.2 Confidence

The results of both experiments provide statistically significant evidence supporting the hypothesis that the use of a credibility-weighted tool rather than an unweighted tool positively affects user confidence. This is consistent with the impact on variance of subcontractor evaluations. Users who are confident in their evaluations are likely to let the evaluations influence their decisions and will therefore vary their decisions.

5.4.2.3 Usefulness

Finally, Experiment II showed that AEC industry practitioners found the credibility-weighted system to be more useful than the unweighted tool in two different settings. This is consistent with the findings in terms of confidence and variation. If the users are confident in the ratings, and thus vary their evaluations more, they can also be expected to find the tool useful.

6 Contributions and suggestions for future work

This chapter summarizes my research contributions and presents suggestions for future work. The chapter begins with a summary of my research in terms of research questions, research methodology and results. Next follows a discussion of my contributions to the state of research in four different fields. The chapter continues by presenting my contributions to industry, before discussing avenues for future work.

6.1 Summary of Research

The adoption rate of electronic commerce has been slow in the construction industry. Rating systems, which enable information sharing between the participants in online communities, have been a major contributing factor to the success of online consumer market places such as eBay. In contrast to consumer electronic marketplaces, the raters in business-to-business communities such as AEC are skilled and connected, necessitating a reputation mechanism that will account for the relationship between the user and the rater.

In the current AEC practice, decision-makers exchange subjective information about subcontractor performance, but the process is inefficient. Furthermore, the existing commercial rating applications in AEC cannot adequately facilitate the sharing of subjective information between industry practitioners. In addition, researchers focusing on construction bidding have not addressed the issue of integrating ratings from multiple sources. Outside construction engineering and management, researchers have proposed several methodologies to rate electronic commerce vendors. However, none of these methodologies can easily be deployed in AEC given they either 1) rely on input parameters that were difficult to measure 2) rely on ad hoc operators, or 3) require large datasets of rating/transaction data for calibration.

In communication research, source credibility theory explicitly investigates the believability of information   Source credibility has been shown to be applicable in commercial settings as well as for the judging of web content, but little research has investigated its applicability in electronic commerce. A rating system based on source credibility has the potential of overcoming all three of the problems associated with other rating mechanisms.

The purpose of this research project is to investigate the fundamental research question:

How can source credibility theory support rating systems in the procurement of AEC services?

To investigate this question it is necessary to find out:

1) *How is it possible to operationalize source credibility to support the calculation of weights that are based on rater identity in an AEC rating tool?*
2) *How can a rating system based on source credibility theory add value in the process of evaluating AEC subcontractors?*

To investigate this question, my major research methodologies have been modeling and experimentation. I operationalized source credibility into two different rating models. These models were tested in two experiments with two separate groups of participants. In the experiments, the performance of a credibility weighted rating tool was performed to that of a standard, unweighted tool.

Both experiments showed with statistical significance that the credibility-weighted models fared better than an unweighted model when it came to predicting rater weights. In the second experiment, I also showed all the factors of the credibility-weighted model to be statistically significant predictors of rater weights. This research project has therefore provided evidence that:

1) The rating tools TrustBuilder I and II are examples of a methodology through which it is possible to operationalize source credibility theory to calculate rater weights.

In addition, both experiments showed with statistical significance that the participants varied their decisions more using a credibility-weighted tool than when using an unweighted tool. The two experiments also provided evidence that the use of the credibility weighted rating tool increases the users' confidence in the ratings. The second experiment also showed that industry practitioners found the credibility-weighted tool to be more useful than the unweighted tool. As a result this research project gives evidence that:

2) A credibility weighted rating tool adds value in the process of evaluating AEC subcontractor by increasing the decision-maker's confidence in the information provided by the rating tool.

6.2 Contributions to state of research

Section 3.3.1 of the Point of Departure chapter identified the opportunities of this research project to contribute to knowledge in four different fields: *AEC electronic commerce, AEC bidding, Rating Mechanisms for Electronic Commerce,* and *Application of Source Credibility theory*. The below discussion analyzes how this project has contributed to knowledge in the four fields.

6.2.1 Contributions to state of research in *AEC electronic commerce*

As stated in the point of departure, there is little research investigating the applicability and added value of rating tools in AEC electronic commerce. To explore this issue, Experiment I did not only test different rating tools, but also included a setting where the user had no ratings available. The results of Experiment 1 showed that both an unweighted rating tool, as well as a credibility weighted tool, performed better, in terms of user behavior as well as opinions, than a tool with where the ratings were absent. I therefore claim that this research project contributes to research in AEC electronic commerce since it *provides evidence that rating tools can add value in AEC electronic bidding.*

The results of Experiment 1 and 2 also showed that credibility weighted tool performed better, both in terms of user behavior and opinions than, an unweighted rating tool. As a result this research project *provides evidence that weighting ratings using source credibility can add value in a rating system supporting AEC e-bidding***.**

Finally, AEC e-commerce is a relative new field of research where, hitherto, most investigations have been in the form of either modeling or case studies. In this research project, experimentation was the primary means of investigation. The fact that the results were meaningful and statically significant enable me to claim that this research project *provides evidence that experimentation can be used by researchers to investigate the applicability and added value of tools that support electronic commerce in AEC.*

6.2.2 Contributions to state of research in *AEC bidding*

Several researchers [2, 5, 7-10, 103] have proposed tools supporting AEC bidding decisions. In previous research (e.g., [3, 5, 7, 103]), there is also ample evidence of the importance for bidding decisions of criteria that can only be measured subjectively by peer industry practitioners. However, the fact that the reliability of the industry practitioners who provide the ratings may vary has generated little interest among researchers. This research project is the first that accounts for the varying reliability of the sources by proposing an AEC bidding tool that weights the ratings depending t on the source of the ratings.

The results from Experiments I and II consistently show that the proposed formalization of source credibility theory predict rater weights better than an unweighted tool. Experiment II also showed that the all the factors of the TrustBuilder II model contribute to prediction rater weights. I therefore claim that this research project *contributes a methodology for the integration of subjective information from multiple AEC practitioners of varying reliability*.

More specifically, I also claim that this research project *contributes a methodology to formalize source credibility to calculate rater weights in AEC, in which the rater weights depend on the user's perception of the credibility of the rater.*

Another aspect of this investigation, which is of interest to researchers in construction engineering and management, is that the results indicate the feasibility of integrating information across organizational borders in AEC.

Furthermore, there is little research in AEC bidding which investigates the added value of source credibility theory as a basis for a rating system. Experiments I and II both showed that decision makers vary their decisions more when using a credibility-weighted than when using an unweighted rating tool. Since the participants in the experiments also were more confident when using the credibility-weighted tool, this research provides evidence that decision makers have more confidence in ratings that has been weighted by credibility more than they do using unweighted ratings. As a result, this research project *provides evidence that source credibility theory can add value in AEC bidding by increasing the user's confidence in the accuracy of the information*.

6.2.2.1 Contributions to state of research in Rating Mechanisms in Electronic Commerce

As shown in the point of departure chapter, there have been numerous research efforts proposing rating mechanisms (e.g., [17, 19, 22, 65, 121, 122]), which support electronic commerce transactions. However, this research project is the first to propose a rating mechanism based on source credibility theory. The results of the experiments show that this study *provides evidence that it is possible to formalize source credibility to support rating mechanisms in electronic commerce.*

Furthermore there is little research in the field of Rating Mechanisms in Electronic Commerce that compares the added value of different rating mechanisms in the context of B2B electronic commerce transactions. AEC e-bidding is one example of B2B electronic commerce transactions. Therefore, at least for certain types of B2B electronic commerce transaction, this research project *provides evidence that a rating system incorporating source credibility theory provides an added value in B2B electronic commerce transactions compared to a standard, unweighted rating mechanism.*

6.2.2.2 Contributions to state of research investigating the Applicability of Source credibility theory

A substantial body of research investigates the applicability of source credibility in commercial as well as online settings. However, there is little research investigating source credibility in an online, as well as commercial, setting (i.e., electronic commerce).

The task of evaluating subcontractors in AEC electronic commerce settings is a situation characterized by 1) substantial benefits from online information sharing, as well as 2) opportunities for deceit. I therefore claim that this research project *provides evidence that source credibility can be applied to construct weights given to information from different sources in an online commercial setting where there are substantial benefits from online information sharing as well as opportunities for deceit.*

The results of Experiment I also indicate that, consistent with previous research, rating discrepancy [39], as well as total rater credibility [38], impact decision makers who aggregate information from multiple sources.

6.3 Contributions to Industry

This research project has several important practical implications in the AEC industry. First of all, it shows that it is possible to construct a rating mechanism that enables the systematic sharing of subjective information across the industry. The results also point to the possibility for an AEC decision maker to take advantage of integrated ratings provided by raters who are known as well as unknown, and who are of varying credibility. This research project also shows that if a rating mechanism takes into account the organizational affiliation of the rater and the user, inter organizational rating systems are feasible in AEC. As a result, Internet based rating systems offer the opportunity for AEC decision makers to pool their knowledge across organizational borders in order to obtain superior information about subcontractor performance. As shown, this is already done in practice to a limited extent, using a manual and informal process. In this project, I have presented a rating system based on source credibility, as a possible methodology to formalize and automate this process.

This research project also gives evidence that the type of rating mechanism used influences bidding-decisions. The results show that users were more confident in the accuracy of the information when the rater weights reflected their personal beliefs. AEC managers should therefore consider implementing personalized rating mechanisms that account for the relationship between the user and the rater. This approach can increase the users' confidence in the ratings and ultimately accelerate the adoption of electronic commerce in the industry.

However, this research project also points to the fact that in a competitive bid environment the potential for subcontractor ratings to directly influence decisions is limited. One conclusion is that the manner in which decision makers evaluate and select subcontractors is a function not only of the available information, but also of the design of the marketplace. Still, as has been pointed out by the participants in the experiments, reliable ratings can also be used after the subcontractors have been selected. General contractors can pro-actively manage subcontractors in order to mitigate potential problems in advance.

6.4 Opportunities for Future Research

This section identifies the major opportunities for future research that this research project brings about in the four research fields of AEC Electronic Commerce, AEC Bidding, Rating Mechanisms for Electronic Commerce, and the Applicability o Source Credibility Theory.

6.4.1 AEC Electronic Commerce

One interesting future field of research is the integration of this project with other work done in AEC electronic commerce. In particular, there is an opportunity to study how a rating system incorporating source credibility could affect trust decisions in Zolin et al's [1] trust model, as well as bid decisions using Tseng and Lin's [2] bid evaluation tool.

Furthermore, if a credibility-weighted tool were to be deployed in a real industry setting, case study research could further investigate the added value of source credibility in AEC electronic commerce.

6.4.2 AEC bidding

To maximize the usefulness of a rating tool to AEC industry practitioners, the tool should incorporate all three of the types of information, which the point of departure chapter discusses; 1) Objective measurements (e.g. project experience), 2)

Subjective measurements provided by a reputable third party (e.g., credit ratings), and 3) Subjective measurements provided by peer industry practitioners. While other researchers (e.g., ([5, 104, 107]) have investigated the integration of the first two types of criteria, this research project has focused on the last type. As a result, there is an opportunity to investigate how all these types of information could be integrated in a rating model that supports AEC bidding decisions. A decision-maker who has access to a subcontractor's project experience, and credit rating, as well as credibility-weighted ratings of its collaborativeness, benefits from an improved chance of making an informed decision. Researchers in construction engineering and management have proposed different methodologies (e.g. MAUT [4], Fuzzy Set [6], AHP [8]) to aggregate ratings into an overall performance indicator. Consequently, there exists an opportunity for researchers to investigate whether it is possible to improve these existing evaluation tools by also incorporating criteria where the information have been weighted by rater credibility.

6.4.3 Rating Mechanisms in Electronic Commerce

When the rating tool proposed in this project estimates rater weights, the calculations are based on direct user input only. As I have shown, there are many other techniques to assign weights in a rating mechanism, such as collaborative filtering, statistical filters, and network of trust. It is likely that the performance of ratings tool can be improved by applying a combination of these different techniques. One could, for example, envision a tool that applied source credibility theory for those cases where the user knows the rater, while applying collaborative filtering to assign weights when the rater is unknown. The development and testing of such a rating tool constitutes an interesting research opportunity.

Another research opportunity is to study decision maker behavior, in the context of real transactions, to further investigate the added value of different rating mechanisms in B2B electronic commerce.

6.4.4 Applicability of Source Credibility Theory

This research project left out several factors that researchers have shown to influence decision-makers who aggregate information. An interesting field of research would be to investigate to what extent factors such as time and message framing impact the weight of a rating mechanism.

Since this research project applied the original McCroskey scale to measure credibility, future research could investigate whether the results would differ when applying other scales (including [33, 36, 37]) to measure the construct. A related area of study is how the different dimensions of source credibility interact with the degree of subjectivism associated with of criteria that is being rated. Does the way users think about credibility differ when the raters are evaluating "maintenance of schedule" rather than the more subjective "ability to cooperate"? One hypothesis, which future research could investigate, is that the importance of trustworthiness relative to expertise increases as the criteria become more subjective.

In the field of applicability of source credibility, future research could also investigate user behavior when an organization is the source of the ratings. In this project, the users evaluated the credibility of "a typical project manager at CalGC" and did not rate the credibility of the organization (CalGC). However, in an industry setting, there may be both legal and practical reasons for listing the organization as the provider of the ratings. In marketing research, several studies (e.g., [27, 170, 171]) have investigated how consumers' perceive the credibility of brand names and companies. In communication research, Newhagen and Nass [40] have shown that people associate the credibility of newspaper articles with the organization (the newspaper) while they consider the anchorpersons to be the source of TV news. Furthermore, in the area of online credibility, studies (e.g., [29, 30]) have evaluated the credibility of websites, which users generally perceive as representing an organization rather than a person. It would therefore be interesting to perform a comparative study in the context of AEC rating systems. Industry practitioners could first evaluate the credibility of a set of contractors and peer practitioners. In the next step, they would evaluate bids based on ratings, which either list the organization or the person as the source.

A final interesting research opportunity would be to, investigate, in a real industry setting, the applicability of source credibility theory operationalized in a credibility-weighted rating mechanism.

7 Bibliography

[1] R. Zolin, “Modeling and Monitoring Trust in Virtual A/E/C Teams,” Stanford University, Stanford, CIFE Working Paper 62, 2000.

[2] H. P. Tseng and P. H. Lin, “An accelerated subcontracting and procurement model for construction projects,” *Automation in Construction*, vol. 11, pp. 105-125, 2002.

[3] G. S. Birrel, “Bid Appraisals Incorporating Quantified Past Performances by Contractors,” presented at AACE Transactions, D.1, Morgantown, 1988.

[4] J. S. Russel, D. E. Hancher, and M. J. Skibniewski, “Contractor prequalification data for construction owners,” *Construction Management and Economics*, vol. 10, pp. 117-135, 1992.

[5] G. D. Holt, P. O. Olomolaiye, and F. C. Harris, “Factors Influencing U.K. Construction Clients' Choice of Contractor,” *Building and Environment*, vol. 29, 1995.

[6] M. I. Okoroth and W. B. Torrance, “A model for subcontractor selection in refurbishment projects,” *Construction Management and Economics*, vol. 17, pp. 315-327, 1998.

[7] G. D. Holt, “Classifying construction contractors,” *Building Research and Information*, vol. 25, 1996.

[8] P. Sik-Wah Fong and S. Kit-Yung Choi, “Final Contractor selection using the analytical hirarchy process,” *Construction Management and Economics*, vol. 18, pp. 547-557, 1999.

[9] M. Wanous, A. H. Boussabaine, and J. Lewis, “To bid or not to bid: a parametric solution,” *Construction Management and Economics*, vol. 18, 1999.

[10] E. Palaneeswaran and M. M. Kumaraswamy, “Contractor Selection for Design/Build Projects,” *Construction Engineering and Management*, vol. 126, 2000.

[11] E. A. Chinyio, P. O. Olomolaiye, S. T. Kometa, and F. C. Harris, “A needs-based methodology for classifying construction clients and selecting contractors,” *Construction Management and Economics*, vol. 16, 1998.

[12] D. Lucking-Reiley, D. Bryan, N. Prasad, and D. Reeves, “Pennies from eBay: the Determinants of Price in Online Auctions,” , Working Paper 2000. <http://w3.arizona.edu/~econ/reiley.html>

[13] S. Dewan and V. Hsu, “Trust in Electronic Markets: Price Discovery in Generalist Versus Specialty Online Auctions,” , Working Paper 2001. <http://databases.si.umich.edu/reputations/bib/papers/Dewan&Hsu.doc>

[14] D. Houser and J. Wooders, “Reputation in Auctions: Theory, and Evidence from eBay,” Department of Economics, University of Arizona, Working Paper 2000.

[15] P. Reznick and R. Zeckhauser, “Trust Among Stangers in Internet Transactions: Empirical Analysis of eBay's Reputation System,” *Journal of Industrial Economics*, vol. Fourthcoming, 2001.

[16] P. Ratnasingham and K. Kulmar, “Trading Partner Trust in Electronic Commerce Participation,” presented at International Conference on Information Systems (ICIS-2000), Brisbane, Australia, 2000.

[17] P. Reznick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, “Grouplens: An open architecture for collaborative filtering of netnews,” presented at Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work, 1994.

[18] M. Chen and J. P. Singh, “Computing and Using Reputations for Internet Ratings,” presented at ACM Conference on Electronic Commerce (EC'01), Tampa, 2001.

[19] C. Dellarocas, “Immunizing online reputation reporting systems against unfair ratings and discriminatory behavior,” presented at Proceedings of the 2nd ACM Conference on Electronic Commerce, Minneapolies, Mn, 2000.

[20] L.-H. Chen and T.-W. Chiou, “A fuzzy credit-rating approach for commercial loans: a Taiwan Case,” *Omega, International Journal of Management*, vol. 27, pp. 407-419, 1999.

[21] R. Khare and A. Rifkin, “Weaving the Web of Trust,” Computer Science, California Institute of Technology, Pasadena, Working Draft 1997. <http://www.cs.caltech.edu/~adam/local/trust.html>

[22] A. Abdul-Rahman and S. Hailes, “Supporting Trust in Virtual Communities,” presented at Hawaii's International Conference on Systems Sciences, Maui, Hawaii, 2000.

[23] J. C. McCroskey and T. J. Young, “Ethos and Credibility: The Construct and its Measurement after three decades,” *Central States Speech Journal*, vol. 32, pp. 24-34, 1981.

[24] C. I. Hovland, I. L. Janis, and H. H. Kelley, *Communication and Persuasion*. New Haven: Yale University Press, 1953.

[25] M. H. Birnhaum and S. E. Stegner, “Source Credibility in Social Judgment: Bias, Expertise and the Judge's point of view,” *Journal of Personality and Social Psychology*, vol. 37, pp. 48-74, 1979.

[26] M. Higgins, “Meta-information ; and time: Factors in human decision making,” *Journal of the American Society For Information Science*, vol. 50, pp. 132-139, 1999.

[27] Vandenbergh, Soley, and Reid, “Factor Study of Advertiser credibility,” *Journalism Quarterly*, vol. 58, 1981.

[28] D. Grewal, J. Gotlieb, and H. Marmorstein, “The Moderating Effects of Message Framing and Source Credibility On the Price-Perceived Risk Relationship,” *Journal of Consumer Research*, vol. 21, pp. 145-153, 1994.

[29] B. Fogg, J. Marshall, O. Laraki, A. Osipovich, C. Varma, N. Fang, J. Paul, A. Rangnekar, J. Shon, P. Swani, and M. Treinen, “What Makes Web Sites credible? A Report on a Large Quantiative Study,” presented at CHI 2001, Seattle, 2001.

[30] W. M. Silberg, G. D. Lundberg, and R. A. Musacchio, “Assessing Controling and Assuring the Quality of Medical Information on the Internet: Caveant Lector et Viewor - Let the Reader and Viewer Beware,” *Journal of the American Medical Association*, vol. 277, pp. 1244-1245, 1997.

[31] R. Critchfield, “Credibility and Web Site Design,” Warner Southern College, Lake Wales 1998. <http://e-

library.ehu.unibel.by/lobko/Common/LearnWeb/Critchfield%20Credibility%20and%20Web%20Site%20Design.htm>

[32] M. E. Rosenbaum and I. P. Levin, "Impression formation as a function of source credibility and order of presentation of contradictory information," *Journal of Personality and Social Psychology*, vol. 10, pp. 167-174, 1968.

[33] K. Andersen and T. Clevenger, Jr., "A Summary of Experimental Research in Ethos," *Speech Monographs*, vol. 30, pp. 59-78, 1963.

[34] R. L. Applbaum and K. W. E. Anatol, "Dimensions of Source Credibility: A Test for Reproductivity," *Speech Monographs*, vol. 40, pp. 230-237, 1973.

[35] J. C. McCroskey, "Scales for the measurement of ethos," *Speech Monographs*, vol. 33, pp. 65-72, 1966.

[36] J. C. McCroskey and J. J. Teven, "Goodwill: A Reexamination of the Construct and Its Measurement," *Communication Monographs*, vol. 66, pp. 90-103, 1999.

[37] D. K. Berlo, J. B. Lemert, and R. Mertz, "Dimensions for evaluating the acceptability of message sources," *Public Opinion Quarterly*, vol. 33, pp. 536-576, 1969.

[38] E. F. Stone and D. L. Stone, "The effects of multiple sources of performance feedback favorability on self-perceived task competence and perceived feedback accuracy," *Journal of Management*, vol. 10, 1984.

[39] E. F. Stone and D. L. Stone, "The effects of feedback consistency and feedback favorability on self-perceived task competence and perceived feedback accuracy," *Organizational Behavior and Human Decision Processes*, vol. 36, pp. 167-185, 1985.

[40] J. Newhagen and C. Nass, "Differential Criteria for Evaluating Credibility of Newspapers and TV News," *Journalims Quarterly*, vol. 66, pp. 277-284, 1989.

[41] B. Fogg and H. Tseng, "The Elements of Computer Credibility," presented at CHI 99, Pittsburgh, 1999.

[42] M. D. Albright and P. E. Levy, "The Effects of Source Credibility and Performance Rating Disrepancy on Reactions to MultipleRaters," *Journal of Applid Social Psychology*, vol. 25, pp. 557-600, 1995.

[43] Trepper, *E-commerce Strategies*: Microsoft Press, 2000.

[44] T. Malone and R. Laubacher, “The Dawn of the E-lance Economy,” *Harvard Business Review*, pp. 145-152, 1998.

[45] D. Lucking-Reiley and D. F. Spulber, “Business-to-Business Electronic Commerce,” , Prepared for the *Journal of Economic Perspectives* 2000.

[46] C. Bass, “Speach At CEE 320, Stanford University,” , 2000

[47] B. Welty and I. Becerra-Fernandez, “Managing Trust and Commitment in Collaborative Supply Chain Relationships,” *Communications of the ACM*, 2001.

[48] W. A. Hanson, *Principles of Internet Marketing*. Cincinnati: South-Western College Publishing, 1999.

[49] R. Drummond, “XML What's still needed for B2B (Industry Trend or Event),” *e-business advisor*, 2000.

[50] J. A. List and D. Lucking-Reiley, “Bidding Behavior and Decision Costs in Field Experiments,” , Working Paper 2000. <http://w3.arizona.edu/~econ/reiley.html>

[51] R. G. Eccles, “The Quasifirm in the construction industry,” *Journal of Economic Behavior and Organization*, vol. 2, pp. 335-337, 1981.

[52] P. Reznick, R. Zeckhauser, E. Friedman, and K. Kuwabara, “Reputation Systems: Facilitating Trust in Internet Interactions,” *Communications of the ACM*, vol. 43, 2000.

[53] C. M. D. Group, “Products and Services: CMD Bulletin.com,” , 2002.<http://www.cmdg.com/products/cmd_bulletincom.html>.

[54] F. Dodge, “Project Center,” , 2002.<http://www.construction.com/ProjectCenter/>.

[55] Caltran, “Bids Opened & Awarded Contracts,” , 2002.<http://www.dot.ca.gov/hq/esc/oe/bidsopened.html>.

[56] R. E. Levitt and N. M. Samelson, *Construction Safety Management*, 2 ed. New York: John Wiley & Sons, 1993.

[57] “The Basics,” in *Managing Engineering and Construction Companies: Class Reader CEE 246*, R. E. Levitt, Ed. Stanford: Stanford Bookstore, 1999.

[58] D. Bradstreet, “Sample Supplier Evaluation Report,” : Telebase Business Research Center, 2000.<http://www.hoovers.telebase.com/ser_smpl.htm>.

[59] M. Ekstrom, J. Taylor, H. Bjornsson, and J. A. Arnold, “Developing Electronic Models for Bidding,” Stanford University, Stanford, CIFE Technical Report 1999.

[60] P. J. Cook, *Bidding for the general contractor*. Kingston: Means Co, 1985.

[61] eBay, “About eBay,” , 2000.<http://pages.ebay.com/community/aboutebay/>.

[62] L. F. Kaiser and M. Kaiser, *The Official eBay Guide*. New York: Simon & Schuster, 1999.

[63] epinon.com, “"@Home, Netscape, Yahoo, Veterans Announce Epinions.com,” , 1999.

[64] L. Guernsey, “Sites Turn Questions and Answers Into a Free-for-All, but Sometimes the Facts Get Trampled,” *New York Times*, 2000.

[65] G. Zacharia, A. Moukas, and P. Maes, “Collaborative Reputation Mechanisms in Electroni Marketplaces,” presented at Thirty-second Annual Hawaii International Conference on
System Sciences (HICSS-32), Wailea, Hawaii, 1999.

[66] Open Ratings, “Implementing Open Ratings - A User Guide to the Open Ratings service,” Open Ratings, Cambridge, MA 2000.

[67] S. B. etat, “Godkjenningskatalogen - Godkjente foretak,” , vol. 2000: Statens Bygningstekniske etat, 2000.<http://www.be.no/FMPro?-DB=foretak&-lay=detail&-format=%2fforetak%2ffinn.html&-View>.

[68] N. M. , “Personal Discussion,” , 2001

[69] eu-supply, “Online Bidding,” : eu-supply, 2002.<http://www.eu-supply.com/service.asp?typ=olb#>.

[70] Buildpoint, “Company Overview,” : Buildpoint, 2002.<http://www.buildpoint.com/company.asp>.

[71] Ratingsource, “How the Subcontractor Monitoring Database Works,” , vol. 2002: Ratingsource, 2002.<http://www.ratingsource.com/brochure/sub_monitor4.htm>.

[72] Struxicon, “Struxicon Raises the Bar for Critical Background Checks in Construction Industry,” , vol. 2000: Struxicon, 2000.<http://www.struxicon.com/info/news/news5-9-00.asp>.

[73] Nexis, “About Nexis.com,” : Nexis, 2000.<http://www.lexis-nexis.com/business/about.htm>.

[74] Coase, “The Nature of the Firm,” , 1937.

[75] O. E. Williamson, *Markets and Hierachies: Analysis and Antritrust Implications*. New York: The Free Press, 1975.

[76] O. E. Williamson, “Comparative Economic Organization: The Analysis of Discrete Structural Alternatives,” *Administrative Science Quarterly*, vol. 36, pp. 269-296, 1991.

[77] S. Winter, “The Boundaries of the Firms:
A System Perspective on the Implications of Information Technology,” : The Warton School, University of Pennsylvania, 2000.<http://www.si.umich.edu/ICOS/Presentations/20000114/>.

[78] S. Gunnarson and R. E. Levitt, “Is a building project a hierarchy or a market? (A Review of Current Literature and Implications for Orgnanizational and Contractural Structure),” presented at 7th Internet Congress, Copenhaguen, 1982.

[79] J. D. Thompson, *Organizations in action*. New York: McGraw-Hill Book Company, 1967.

[80] S. E. Masden, J. W. M. Jr, and E. A. Snyder, “The Costs of Organization,” *Journal of Law ,Economics and Organization*, vol. 4, pp. 181-198, 1991.

[81] O. E. Williamson, “Comparative Economic Organization: The Analysis of Discrete Structural Alternatives",” *Administrative Science Quarterly*, vol. 36, pp. 269-296, 1991``.

[82] K. Kim, B. C. P. Jr., and C. J. P. Jr., “Agent-Based Electronic Markets for Supply Chain Coordination,” presented at KBEM'00Knowledge-based Electronic Markets, Austion, 2000.

[83] H. Bjornsson and J. Taylor, “Construction Supply Chain Improvements through Internet Pooled Procurement,” presented at Seventh Annual

Conference of the International Group for Lean Construction (IGLC-7), Berkeley, 1999.

[84] D. De la Hoz Sanchez and F. Ballester Munuz, “Internet Strategies. It's time for the silent Internet,” presented at ECCE ICT Symposium 2001, Espoo, Finland, 2001.

[85] A. M. Clark, B. L. Atkin, M. P. Betts, and D. A. Smith, “Benchmarking the use of IT to support supplier management in construction,” *Electronic Journal of*

Information Technology in Construction, vol. 4, 1999.

[86] T. Koivu, “The Future of Product Modeling and Interoperability in the AEC/FM Industry: A workshop in Budapest, Nov 30, 2001,” CIFE, Stanford, Stanford, CIFE, Working Paper 2001.

[87] A. Barron and M. Fischer, “Potential Benefits of Internet-Based Project Control Systems - A study on Monthly Billings Processing,” CIFE, Stanford University, Stanford, CIFE, Technical Report 127, 2001. <http://www.stanford.edu/group/CIFE/Publications/index.html>

[88] B.-C. Björk, “Document Management - A key IT technology for the construction industry,” presented at ECCE ICT Symposium 2001, Espoo, Finland, 2001.

[89] R. Howard and E. Petersen, “Monitoring Communications in Partnering projects,” *Electronic Journal of Information Technology in Construction*, vol. 6, 2001.

[90] M. Suchocki, “Successfully Adopting Collaborative Technology,” presented at ECCE ICT Symposium 2001, Espoo, Finland, 2001.

[91] M. Murray and A. Lai, “The Management of Construction Projects Using Web Sites,” presented at ECCE ICT Symposium 2001, Espoo, Finland, 2001.

[92] M. Mortensen, “The Effects of Project Management Software on Project-Wide Communication Patterns,” Stanford, Stanford, CIFE Working Paper 1999.

[93] W. Behrman and B. Paulsson, “Best Practices for the Development and Use of XML Data Interchange Standards,” , vol. 2002: CIFE, Stanford University, 2000.<http://www.stanford.edu/group/CIFE/Research/index_bd.html>.

[94] F. Tolman, J. Stephens, R. Steinmann, R. v. Rees, M. Böhms, and A. Zarli, "bcXML, an XML Vocabulary for Building and Construction," presented at ECCE ICT Symposium 2001, Espoo, Finland, 2001.

[95] A. Pouria and T. Froese, "Transaction and implementation standards in AEC/FM industry," , Victoria, BC., 2001.

[96] M. Shreyer and J. Schwarte, "New Building Materials - Knowledge Transfer via the World Wide Web," presented at ECCE ICT Symposium 2001, Espoo, Finland, 2001.

[97] A. a. R. Zarli, O., "Requirements and Technology Integration for IT-Based Business-Oriented Frameworks in Building and Construction," *Electronic Journal*

of Information Technology in Construction, vol. 4, 1999.

[98] J. A. Arnold, "Information Interoperation for Internet-based Engineering Analysis," in *Civil Engineering*: Stanforde University, 2000

[99] C. D. Leonard, "ICT at work for the LSE Procurement," , vol. 2002: PROCURE, 2000.<http://cic.vtt.fi/projects/procure/public.html>.

[100] W. Park, *Construction Bidding for Profit*. New York: John Wiley & Sons, 1979.

[101] G. Runeson and M. Skitmore, "Tendering Theory Revisited," *Construction Management and Economics*, vol. 17, pp. 285-296, 1997.

[102] J. S. Russell, D. E. Hancher, and M. J. Skibniewski, "Contractor prequalification data for construction owners," *Construction Management and Economics*, vol. 10, 1992.

[103] D. Drew and M. Skitmore, "The effect of contract type and size on competiveness in bidding," *Construction Management and Economics*, vol. 15`, pp. 469-489, 1997.

[104] S. Kale and D. Arditi, "General Contractor's relationhips with subcontractors: a strategic asset," *Construction Management and Economics*, vol. 19, pp. 541-549, 2001.

[105] T. F. J. Gilbane, "Subcontractor Bidding Strategy," in *Civil Engineering*. Boston: Massachusetts Institute of Technology, 1975

[106] A. A. Shash, “Subcontractor's Bidding Decisions,” *Journal of Construction Engineering and Management*, 1998.

[107] L. Friedman, “A competitive bidding strategy,” *Operations Research*, vol. 4, 1956.

[108] M. Gates, “Bidding strategies and probabilities,” *Journal of Construction Engineering and Management*, vol. 119, pp. 131-147, 1967.

[109] D. von Winterfeld and W. Edwards, *Decision Analysis and Behavioral Research*. Cambridge, England: Cambridge University Press, 1986.

[110] P. E. D. Love, M. Skitmore, and G. Earl, “Selecting a suitable procurement method for a building project,” *Construction Management and Economics*, vol. 16, pp. 221-233, 1998.

[111] T. L. Saaty, *The Analytic Hierarchy Process, Planning, Priority, and Resource Allocation*. New York: McGraw-Hill, 1980.

[112] J. H. M. Tah and V. Carr, “A proposal for construction project risk assessment using fuzzy locic,” *Construction Management and Economics*, vol. 18, 2000.

[113] A. M. Elazouni and F. G. Metwally, “D-SUB: Decision Support System for Subcontracting Construction Works,” *Journal of Construction Engineering and Management*, vol. 126, 2000.

[114] C. Dellarocas, “Analyzing the efficiency of eBay-like online reputation reporting mechanisms,” Center for eBusiness@MIT, Working Paper 2001.

[115] K. Swearingen and R. Siinha, “*Beyond Algorithms: An HCI Perspective on Recommender Systems*,” presented at ACM SIGIR Workshop on Recommender Systems Sept 2001, 2001.

[116] S. Milgram, “The Small World Problem,” *Psychology Today*, pp. 60-76, 1967.

[117] J. Guare, *Six Degrees of Separation*: Random House, 1990.

[118] B. Yu, M. Venkatraman, and M. P. Singh, “An adaptive social network for information access: Theoretical and experimental reuslts,” *Applied Artificial Intelligence*, 2000.

[119] A. R. Rohit Khare, “Weaving the Web of Trust,” Computer Science, California Institute of Technology, Pasadena, Working Draft 1997. <http://www.cs.caltech.edu/~adam/local/trust.html>

[120] C. Ono, D. Kanetom, K. Kim, J. Boyd C. Paulson, M. Cutkosky, and J. Charles J. Petrie, “Trust-based Facilitator for e-Partnership,” *Unpublished*, 2000.

[121] U. Shardanand and P. Maes, “Social Information Filtering: Algorithms for Automating "Word of Mouth",” presented at ACM Special Interest Group on Computer-Human Interaction, 1995.

[122] P. Melville, R. J. Mooney, and R. Nagarajan, “Content-Boosted Collaborative Filtering,” presented at 2001 ACM SIGIR Workshop on Recommender Systems, 2001.

[123] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, “Item-based Collaborative Filtering Recommendation Algorithms,” presented at WWW10, Hong Kong, 2001.

[124] S. Aguzzoli, P. Avesani, and P. Massa, “Compositional Recommender Systems Using Case-Based Raeasoning Approach,” presented at 2001 ACM SIGIR Workshop on Recommender Systems, 2001.

[125] C. Avery, Reznick, P., Zeckhauser, R., “The Market for Evaluations,” *American Economic Review*, vol. 89, pp. 564-584, 1999.

[126] W. Siler, “Constructing Fuzzy,” , 2000.<http://members.aol.com/wsiler/>.

[127] L. A. Zadeh, “Fuzzy sets,” *Inf. Control*, vol. 8, pp. 338-353, 1965.

[128] H.-J. Zimmerman and P. Zysno, “Decision Evaluations by hierarchical aggregation of information,” *Fuzzy Sets and Systems*, vol. 10, pp. 243-260, 1983.

[129] S. G. Romaniuk and L. O. Hall, “Decision making on creditworthiness using a fuzzy connectionist model,” *Fuzzy Sets and Systems*, vol. 15-22, 1992.

[130] D. Gambetta, “Can We Trust?,” in *Trust: Making and Breaking of Cooperative Relations*, D. Gambetta, Ed. Oxford: Blackwell, 1990.

[131] R. A. a. J. E. M. Howard, “Readings on the Principles and Applications of Decision Analysis,” . Menlo Park: Strategic Decisions Group, 1984

[132] C. Castelfranchi and R. Falcone, “Trust Is Much More than Subjective Probability: Mental Components and Sources of Trust,” presented at Hawaii International Conference on System Sciences, Maui, Hawaii, 1998.

[133] H.-J. Zimmerman, *Fuzzy Set Theory-And Its Applications*. Norwell: Kluwer, 1996.

[134] L. Cooper, *The Rhetoric of Aristotle*. New York: Appleton-Century-Crofts, 1932.

[135] Aristotle, *On Rhetoric: A Theory of Civic Discource*. New York: Oxford University Press, 1991.

[136] C. I. Hovland and W. Weiss, “The Influence of Source Credibility on Communication Effectiveness,” *Public Opinion Quarterly*, 1951.

[137] C. D. Fisher, D. R. Ilgen, and W. D. Hover, “Source Credibility, information favorability, and job offer acceptance,” *Academy of Management Journal*, vol. 22, pp. 94-103, 1979.

[138] R. R. Harmon and K. A. Coney, “The persuasive effects of source credibility in buy and lease situations,” *Journal of Marketing Research*, vol. 19, pp. 255-260, 1982.

[139] H. Constantinides and J. Swenson, “Credibility and Medical Web Sites: A Literature Review,” Departmeny of Rhetoric: University of Minnesota, St Paul 2000.

[140] American Medical Association, “Guidelins for Medical and Health Information Sites on the Internet: Principles Governing AMA WebSites,” : AMA, 2001.<http://www.ama-assn.org/ama/pub/category/3959.html>.

[141] S. Goodwin, K. Ovnic, and H. Korschun, “'Report Card' for health-related web sites,” in *Emery Healsth Sciences Press Release*, 1999.<http://www.emory.edu/WHSC/HSNEWS/releases/aug99/080999web.html>.

[142] Healthwise, “Web Content,” : Healthwise, 2002.<http://www.healthwise.org/p_web-content.html>.

[143] M. S. Eastin, “Credibility Assessments of Online Health Information: The Effects of Source Expertise and Knowledge of Content,” *Journal of Computer-Mediated Communication*, 2001.

[144] A. J. Flanagin and M. J. Metzger, “Perceptions of Internet information credibility,” *Journalism and Mass Communication Quarterly*, vol. 77, pp. 515-540, 2000.

[145] E. F. Stone and D. L. Stone, “The effects of multiple sources of performance feedback and feedback favorability on self perceived task competence and perceived feedback accuracy,” *Journal of Management*, vol. 10, pp. 371-378, 1984.

[146] D. M. Mackie, L. T. Worth, and A. G. Ascuncion, “Processing of Persuasive In-group Messages,” *Journal of Personality and Social Psychology*, vol. 58, pp. 812-822, 1990.

[147] N. H. Andersson, “Cognitive Theory of Judgment and Decision,” in *Contributions to Information Integration Theory*, vol. 1, N. H. Andersson, Ed. Hillsdal: Lawrence Elbaum Associates, 1991.

[148] J. A. Rice, *Mathematical Statistics and Data Analysis*. Belmont: Wadsworth Publishing Company, 1995.

[149] A. C. Guenther, “Extremity of attitude and trust in media news coverage of issues,” in *Communication*. Stanford: Stanford University, 1989

[150] M. W. Singletary, “Accuracy in news reporting: A review of the research. *ANPA News Research Report*,” *Journalism Quarterly*, vol. 53, pp. 316-319, 1976.

[151] R. B. Rubin, P. Palmgreen, and H. E. Sypher, “Communication Research Measures: A Sourcebook,” . New York: Guilford Publications, 1994

[152] T. L. Saaty, *The Analytic Hirarchy Process*. New York: McGraw-Hill, 1980.

[153] C.-L. Hwang and K. Yoon, *Multiple Attribute Decision Making - Methods and Applications*. New York: Springer-Verlag, 1987.

[154] J. S. Uebersax, “Raw Agreement Indices,” , vol. 2001, 2001.<http://ourworld.compuserve.com/homepages/jsuebersax/raw.htm>.

[155] C. C. Clogg, “Latent class models,” in *Handbook of statistical modeling for the social and behavioral sciences*, C. C. C. G. Arminger, & M. E. Sobel, Ed. New York: Plenum, 1995, pp. 311-359.

[156] A. Agresti, “ Modelling patterns of agreement and disagreement,” *Statistical Methods in Medical Research*, pp. 201--218, 1992.

[157] U. Olsson, “Maximum likelihood estimation of the polychoric correlation coefficient,” *Psychometrika*, vol. 44, pp. 443-460, 1979.

[158] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and Psychological Measurement*, vol. 20, pp. 37-46, 1960.

[159] J. S. Uebersax, “Statistical Methods for Rater Agreement,” , vol. 2001, 2001.<http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm>.

[160] Mathworks, “Model Browser User's Guide,” , vol. 2002: Mathworks, 2001.<http://www.mathworks.com/access/helpdesk/help/toolbox/mbc/model/techdo20.shtml>.

[161] J. Thomsen, “The Virtual Design Team: A Proposed Trajectory of Validation Experiments for Computational Emulation Models of Organizations,” CIFE, Stanford, CIFE Working Paper WP 047,.

[162] M. B. Salwen and D. W. Stacks, *An Integrated Approach to Communication Theory and Research*. MahWah: Lawrence Erlbaum Associates Publishers, 1996.

[163] M. J. Clayton, J. C. Kunz, and M. A. Fischer, “The Charette Test Method,” CIFE, Stanford, Techical Report 120, 1998.

[164] R. Mulye, “An empirical comparison of three variants of the AHP and two variants of conjoint analysis,” *Journal of Behavioral Decision Making*, vol. 11, pp. 263-280, 1998.

[165] G. Wright and F. Bolger, “Expertise and Decision Support,” . New York: Plenum Press, 1992

[166] L. Cohen and M. Holliday, *Statistics for social scientists*. London: Harper & Row Ltd, 1982.

[167] R. M. Sirkin, *Statistics for the social sciences*. Thousand Oaks: SAGE Publications, 1995.

[168] W. N. Venables and B. D. Ripley, *Modern Applied Statstics with S-Plus*, 3 ed. New York: Springer, 1999.

[169] H. Sattler and K. Schrim, “Credibility of Product-Preannouncements,” *Fourthcoming in Journal of the Academy of Marketing Science*, 2000.

[170] R. E. G. B.A. Lafferty, “Corporate Credibility's Role in Consumers' Attitudes and Purchase Intentions When a High versus a Low Credibility Endorser Is Used in the Ad,” *Journal of Business Research*, vol. 44, 1999.