Forthcoming in *CIB W65 Symposium September 2002*, Cincinnati, Ohio, USA

A rating system for AEC e-bidding that accounts for rater credibility

Martin A Ekström, Ph.D. Candidate, Stanford University,
mekstrom@stanford.edu
Hans C Björnsson, Professor, Stanford University,
hansbj@stanford.edu

INTRODUCTION

Until recently e-commerce providers asserted how they would " revolutionize the way construction professionals conduct their business".[1] In April 2000 approximately $1 Billion had been invested in around 200 Architecture Engineering Construction (AEC) industry electronic commerce start-ups. In reality, the amount of transactions conducted in these new electronic market places turned out to be very low, and only a handful of them were still in business in February 2002. One possible cause of the AEC industry's slow adoption of electronic commerce is the lack of trust in the participants in electronic marketplaces. Electronic markets facilitates the search for new business partners, but AEC General Contractors, subcontractors, and suppliers are unlikely to do business with firms that are unknown to them. One solution to this problem is the creation of a rating system that gathers and displays information about the past performance of the market participants. The most well-known e-commerce rating system belongs to eBay, the largest and most successful Internet Auction provider, enabling transactions between private parties. eBay currently (Feb 2002) has over 30 million registered users. eBay's users buy and sell items that are varied, and often difficult to describe as well as evaluate. Trust is therefore a prerequisite for the transactions on eBay to take place. To foster trust, eBay has created a system by which the market participants rate each other after each transaction. eBay's success makes it interesting to investigate whether a rating system could support trust also in AEC e-markets. The eBay rating system is simple and intuitive to use. The problem is that all ratings weigh the same, implying that the system is built on the assumption that all raters are equally credible and trustworthy. This assumption may be valid in C2C auctions but we doubt that it applies to complex B2B transactions, such as AEC bidding. More sophisticated electronic commerce rating solutions have been developed incorporating features such as multiple criteria (e.g., Bizrate), network of trust (e.g., epinions.com) and sophisticated statistical analysis (openratings.com). However, we argue that none of these solutions directly addresses the problems facing an AEC electronic market, being either overly simple (ebay and Bizrate), or

[1] See for example About Struxicon at the "CIFE e-commerce in design and construction" summit in March, 2000: http://www.stanford.edu/group/CIFE/ecom.company.summary.html

requiring a substantial amount of transaction data to calculate the ratings (Openratings). We therefore argue that a new approach is necessary. This paper investigates how different theoretical frameworks can be used as a basis for an AEC rating system. We propose a model for calculating ratings of AEC subcontractors based on the user's assessment of rater credibility. The paper reports the results of an experiment where the participants evaluated subcontractors using a both a credibility-weighted rating model and a standard unweighted model. The experiment showed that the credibility-weighted model fared better in terms of user behaviour as well as attitudes.

THEORETICAL FOUNDATIONS OF THE STUDY

There are several candidate methods that can be used for synthesizing uncertain information from sources of varying reliability. Zacharia et al (1999) have proposed a complex rating mechanism where the weight of the ratings depend on the user's trust in the rater along with the rater's previous rating behavior. However this rating mechanism equates trust in a business partner with trust in a rater. It is far from certain that a general contractor that is regarded as a trustworthy employer of subcontractors is also a reliable rater. Zimmerman and Zysno (1983) and others (Chen and Chiou 1999; Romaniuk and Hall 1992)) has applied fuzzy set theory to the credit rating processes conducted by banks. A process which is similar to an AEC subcontractor rating problem. Another potential solution is to deploy subjective probabilities using, for example, Howard's 5-step interview process (Howard 1984) to estimate the expected value and variance associated with a set of subcontractor ratings. A common problem with the methodologies discussed above is that they work very well once one has obtained the right input information. If one models and accurately measures the reliability of a source and how it interacts with the message (rating), the synthesizing of the information can be done in a consistent and non-arbitrary manner. However, without the correct input a synthesizing function provide little value in itself. Instead, we argue that the major problem in creating a functioning rating mechanism seems to be to accurately model and measure the reliability of the sources. Unless this can be done in a non-arbitrary manner, the algorithms serving to calculate the overall rating add little value in a real world situation. A very different way to approach this problem is to take recourse in communication research where source credibility theory has been developed to explicitly judge the credibility of media sources. In the following paragraphs we investigate how source credibility can be operationalized and extended in order to "rate the rater" in an AEC rating system.

Source Credibility theory

The foundations to source credibility theory were laid by Hovland et al (1953) who identified perceived trustworthiness and perceived expertise as the main dimensions of a source's credibility. The higher the trustworthiness and expertise a source is perceived to have, the higher will be the importance given to information coming from that source. Early work in the field applied source credibility in the

context of public opinion (Berlo et al. 1969; Hovland et al. 1953; Hovland and Weiss 1951) and interpersonal communication. (Berlo et al. 1969). Later studies have shown that source credibility theory applies also applies to commercial settings where the receiver of information is evaluating possible transactions (e.g., (Birnhaum and Stegner 1979), (Fisher et al. 1979), (Harmon and Coney 1982)). We hypothesize that source credibility can also be used to assign weights to different raters in an AEC rating system. To operationalize source credibility we propose applying the validated McCroskey (1966) 12 item Likkert scale. The McCroscey scale measures a source's credibility in terms of two factors: Authorativeness (corresponding to Expertise) and Character (Trustworthiness).

However, rater credibility is not the only factor that influences the weight of a rating. .Zacharia et al (1999) point out that time is another important factor. Based on interviews we have found that ratings that were more than two years old were often regarded as substantially less credible than recent ones, even though the discount factors seemed to be highly individualized.

Stone and Stone (1984) reported that information from two sources was perceived to be more credible than information from a single source. As a result, a user evaluating subcontractors would find it useful to know the total rater credibility or the sum of the credibility of all raters that have rated a given subcontractor. Has only one rater with low credibility rated a subcontractor, or is the overall subcontractor rating based on the ratings from several very credible raters?

Feedback consistency has also been found (Stone and Stone 1985; Albright and Levy 1995) to be an important indicator when assessing the accurateness of performance feedback from multiple sources.

TRUSTBUILDER: A RATING TOOL OPERATIONALIZING SOURCE CREDIBILITY THEORY

In order to investigate in how source credibility theory can support AEC ratings the research team designed TrustBuilder, a rating tool that weighs ratings by rater credibility. TrustBuilder uses two types of information that can support the evaluation of rater credibility: direct knowledge about the rater, and knowledge about the rater's organization. This credibility weighted rating tool follow a 3-step process:

Step 1: Credibility Input – TrustBuilder lets the user rate three different types of raters on the 12 item McCroskey credibility scale[2]. The three types of raters being assessed are:

i) Unknown Rater: The user does neither know the rater nor the organization the rater works for. This assessment is used as a baseline measure of credibility**.**

ii) Unknown rater, Known organization: This measure is used to estimate the credibility of the organization the rater is working for.

[2] 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"

iii) Known Rater, Known organization: This measure refers to raters that the user knows personally.

Step 2: Calculation of Rater Weights – The next step is to convert the ratings of rater credibility to weights. TrustBuilder does this through logistic regression in combination with a methodology of pairwise comparisons. Pairwise comparisons have been widely adopted in decision-making tools that apply the Analytic Hierarchy Process (AHP) (Saaty 1980). In TrustBuilder, the users are shown a user interface where a painting subcontractor ("PaintA") has been rated by two raters.

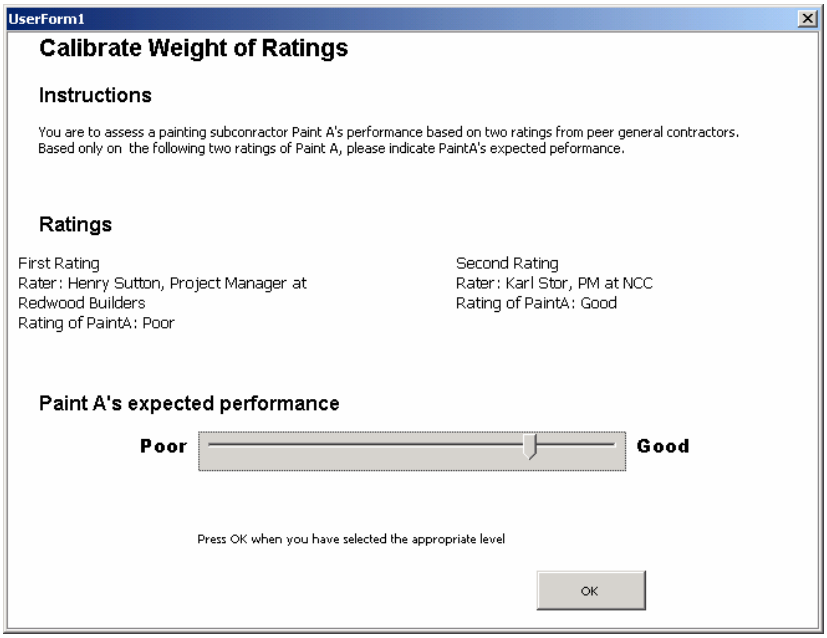

Figure 1: The user performs pair-wise comparison to enable TrustBuilder to calculate rater weights of ratings through pairwise comparisons and logistic regression.

Rater 1 rated PaintA's performance as "Good", while Rater 2 rated it to be "Poor". TrustBuilder asks the users to submit their assessment of PaintA's performance by dragging a continuous slide-bar in between the values "Poor" and "Good". This value (w_{12}) - corresponding to the weight that the user attributes to Rater 1's ratings vis-à-vis Rater 2's - is then used as the target function in a logistic regression to estimate the credibility (C_{ij}) of each rater j from user i's perspective. An overall rating (R_{ik}) of a subcontractor k, customized for user i can then be calculated using the following straight forward formula:

$$R_{ik} = \sum_j R_{ij} * C_{ij} / \sum_j C_{ij} \qquad (1.1)$$

$$R_{jk} \neq 0$$

Furthermore, TrustBuilder uses an adoption of a raw agreement index (see for example (Uebersax 2001)) to calculate rater agreement. The adoption consists of incorporating the notion of rater credibility when calculating rater agreement. Total credibility is simply calculated as the sum of all the raters' individual credibility.

Step 3: Display Ratings and Rater Information – In the prototype user interface the user can see the calculated values for two different subcontractors (see Figure 2.) The user can see the overall rating (weighted by credibility) both on a continuous scale and as a symbolic value. For each subcontractor, she can also see the rater agreement along with the total credibility of all the raters. TrustBuilder also shows the credibility for each rater customized by user preferences. The prototype allows the user to input contingency for each bid and select the best bidder.

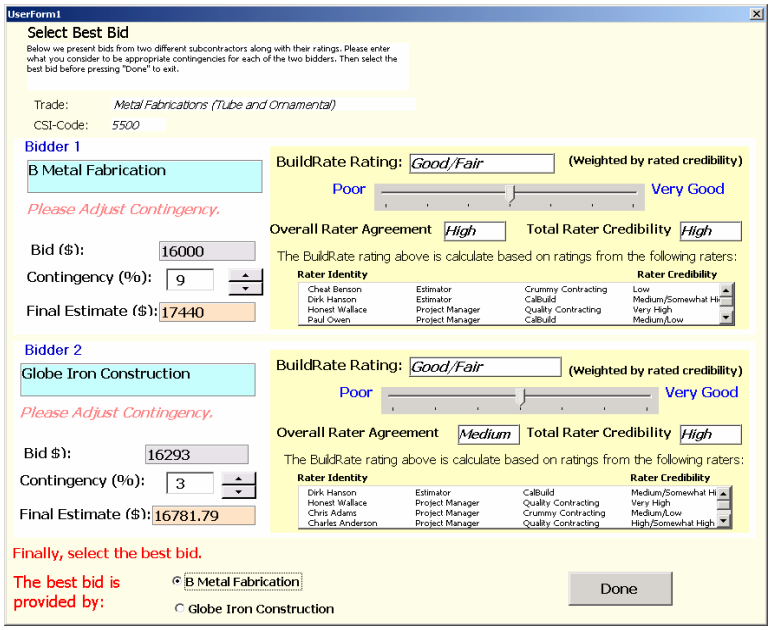

Figure 2: User interface showing bids and ratings for two subcontractors. TrustBuilder also shows rater credibility and agreement. The user inputs contingency for each bid and selects the best bid.

AN EXPERIMENT TO EVALUATE RATING SYSTEMS FOR BIDDING

In order to test the applicability of source credibility theory as a basis for an AEC rating system we designed an experiment that compared two different rating mechanisms along with the absence of ratings:
Unweighted Ratings: The subcontractor's overall ratings are calculated as the average rating where all raters are weighted the same. This is the standard mechanism similar to for example eBay's and Bizrate's systems.
Credibility-weighted Ratings: This tool corresponds to TrustBuilder, described above, where each subcontractor rating is weighted by the user-defined credibility of the rater. The rater is shown the overall rating along with total rater credibility and agreement. The major purpose of the experiment was to compare the performance of this mechanism to that of the unweighted mechanism.
No ratings: The users do not have any ratings to support the subcontractor evaluation. We included this mechanism to have a baseline measure to which we could compare the two other rating mechanisms.

The participants used the three rating mechanisms (or tools) to evaluate subcontractors bidding for the trades subcontracted in the construction of a recently completed $5M office building in San Francisco, California.
The experiment was designed to investigate the following hypotheses:

H1*: In the context of participants making pair wise comparisons, credibility measures based on the McCroskey scale are better than a unweighted (constant) model at predicting the relative weights that users attribute to different raters.*
H2*: Users will vary the contingency added to the subcontractors' bids more when using the credibility-weighted tool than when using the unweighted tool.*
H3*: The use of a credibility-weighted rather than an unweighted tool results in increased user confidence in the user's judgments of overall performance.*
H4*: The aggregated rating of a subcontractor is negatively correlated to the contingency added to the subcontractor's bid.*
H5*: The rater agreement regarding the performance of a subcontractor is negatively correlated to the contingency added to the subcontractor's bid..*
H6*: The total credibility of all the raters that have rated a subcontractor is negatively correlated to the contingency added to the subcontractor's bid.*

Method

Participants

The participants were 16 construction management students, faculty and professionals (ages ranged between 24 and 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). The participants were randomly assigned to the condition regarding the order in which the different rating tools were presented.

Procedure

The experiment was a within subject design that was carried out on an individual basis with each participant being supervised by an instructor. The experiment was carried out on a personal computer at the participant's place of work. The instructor began by showing the participant a 10-minute tutorial introducing the concept of rating systems. The user then started a Microsoft Excel/Visual Basic application that ran the experiment. The participants first calculated rater weights following TrustBuilder procedure outlined above. They then evaluated one pair of subcontractor bidding for each of the project's 17 trades. For each pair of subcontractors, one of the three tools displayed bids, subcontractor ratings and rater information. The participants were asked to add contingency to each bid along with selecting the best bidder.

Manipulation

Given the exploratory nature of the study the experiment was carried out using hypothetical subcontractor ratings and bids.
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 that had rated the subcontractor. However he did not know who had rated the subcontractor.
The tool without ratings was very simple. It consisted of the two subcontractors' names and bids.

Measures

Rater credibility was measured with the McCroskey 12 item credibility scale.
Goodness of fit of model was measured using the sum of squared errors in the pairwise comparisons for the two models.
Bid contingency was measured with a single item. The users entered a number between 0-100% for bid contingency. The contingency was intended to reflect both the user's assessment of the risk buffer that should be added to a bid, as well as the extra cost of managing an under-performing subcontractor.
Users' confidence in their assessments was measured with a single item: "How confident are you in your judgment?"

Results

Goodness of Fit

To test H1 (that a credibility-weighted model was better than an unweighted model at predicting rater weights) a generalized maximum likelihood ratio test was performed (see for example (Rice 1995)). This test compared the sum of squared errors of the unweighted and the credibility weighted models, taking into account the higher degrees of freedom of the credibility-weighted model. The sum of the

errors were considerable larger for the unweighted model (17.52) than for the credibility-weighted model (5.11, $p<0.0001$) (see Figure 3.)

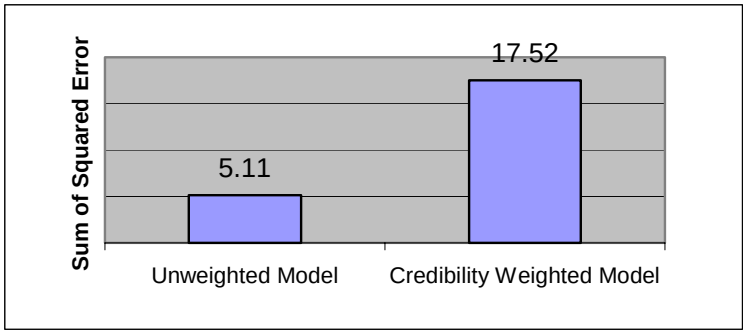

Figure 3: Sum of Squared Error in pairwise comparison using unweighted and credibility-weighted models (N=336). The results are consistent with H1 (that a credibility weighted model is better than an unweighted model at predicting rater weights).

The result indicates that the credibility-weighted model is superior to the unweighted model when it comes to predicting of rater weights (H1). One conclusion is that the TrustBuilder operationalization of source credibility theory can be used to model weights in the context of AEC bidding. This property is a prerequisite for the application of a successful rating tool.
To further test the applicability of source credibility to an AEC bidding context, we performed a principal component factor analysis. The factor analysis showed that the McCroskey scale did indeed factor into the two factors Character and Authoritativeness, indicating that measuring credibility in terms of Expertise and Trustworthiness seem to be valid also in AEC bidding.

Contingency

We used the Sign Test to test whether users varied the contingency more using the credibility-weighted tool (H2. First, we calculated the variance of the contingency assigned by each user when using each of the three tools. The form of data yielded from the survey was deemed suitable for analysis by the Sign Test following the criteria and procedures set out in Cohen and Holliday (1982). As shown in Figure 4, 14 out of 16 users varied the contingency more when using the credibility-weighted tool than when using the unweighted tool (Sign Test: n+ 14, n- 2, $p<0.005$.), users varied their decisions more when using a credibility-weighted tool than when using no tool. (Sign Test: n+ 15, n- 1, $p<0.001$).

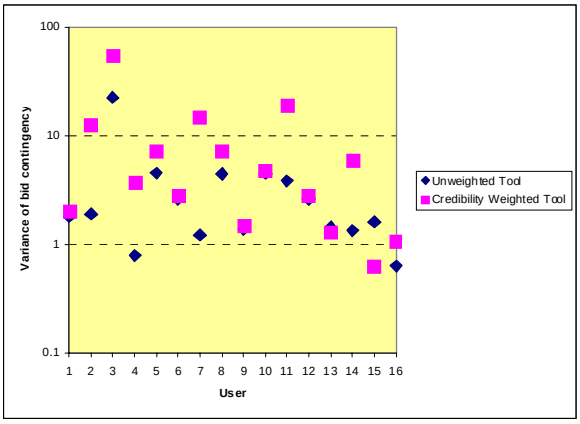

Figure 4: Variance of contingency for each user with the unweighed and credibility weighted tools. 14 out of 16 users had a higher variance when using the credibility weighted tool.

The results indicate that the users will vary their evaluations of subcontractors more using a credibility-weighted tool. As a result, the bidding price will be of less importance. A user would then be less likely to select the lowest bidder when using the credibility-weighted tool than when using the unweighted tool. This is an important finding since the purpose of a decision support tool such as a rating system is to provide the user with information that she trusts enough to act upon.

Confidence

The participants expressed higher confidence in their evaluatons (H2) 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 consistent with the results for the behavioural measure of bid contingency. A confident user will be more likely to vary his decision depending on the information provided by the rating tool.

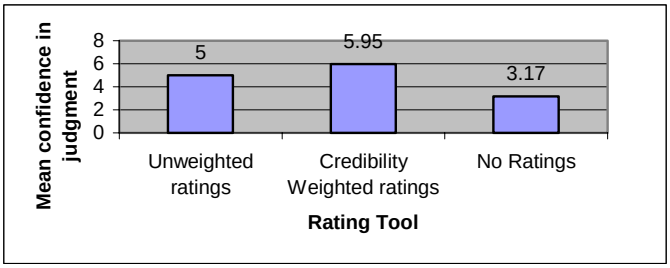

Figure 4: Confidence in judgment using three different rating tools. Users expressed more confidence when using the Credibility weighted tool.

Agreement and Total Credibility

The study also intended to investigate to what extent Average Rating Agreement and Total Credibility influenced bidding decisions (H4-H6). For the tested set of raters, both agreement ($p<0.01$), total credibility ($p<0.05$), and Final (Aggregated) Rating ($p<0.01$)) were shown to influence bid contingency (Figure 5.) The coefficients are all negative since the better the overall rating, the higher agreement of the raters, and the more trustworthy these raters are perceived to be, the lesser amount of bid contingency a user will feel inclined to add.

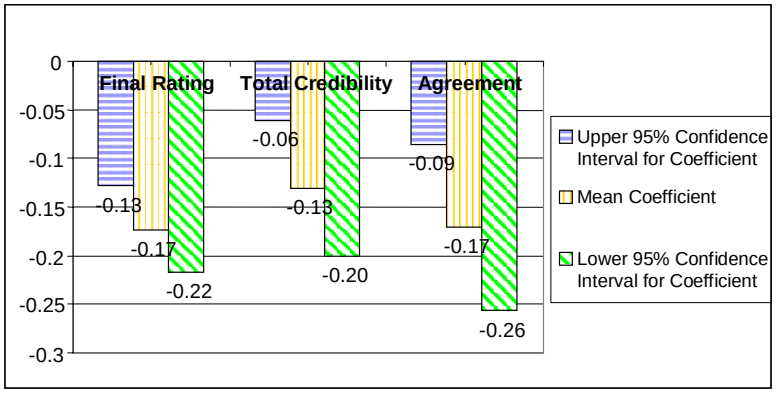

Figure 5: Coefficients in Linear Regression of Bid Contingency using credibility-weighted tool. The results show that the users base their bid contingency decisions not only on the Final Rating, but also on Agreement and Total Credibility.

DISCUSSION

The study suggests that source credibility is a promising basis for an AEC rating system. The experiment found differences between a credibility-weighted and an unweighted model both in terms of a behavioral measure (contingency) and an attitudinal measure (confidence). The results indicate that, by knowing that the ratings are filtered by rater credibility, users become more confident in their decisions, and hence more likely to let the rating tool influence their decisions. That the users trust the ratings is a key aspect for a successful deployment of a rating system in an AEC electronic marketplace. General Contractors that trust the sucontractor ratings provided by a rating system would be less reluctant to hire a subcontractor they had not hired before.

Furthermore the study indicates that especially rater agreement but also total rater credibility affect user's evaluations of subcontractors. If further aggregation is strived for, a rating tool could calculate a final overall rating incorporating the overall credibility weighted rating, the level of agreement, as well as total rater credibility. For a tool that is designed to support human decision-making this level of aggregation may not be called for. However, a credibility-weighted rating

mechanism could also support the automaton of low risk labor-intensive tasks. One example would be an automated bid invitation tool that invites all the subcontractors who fulfil a set of criteria covering (e.g., minimum overall rating, minimum safety record, maximum bond rate, etc) to bid on a job

In future research we plan to repeat the experiment in more realistic conditions, having industry practioners, that are experts in evaluating subcontractor, apply different tools to evaluate a real set of subcontractors. The TrustBuilder rating tool can also be elaborated by refining the 3-step process calculating credibility weights and by extending it to include more than one rated criteria.

Finally, we want to emphasise that a rating system that accounts for rater credibility will not be able to enforce truthful behaviour in an Internet rating system. A prerequisite for a functioning rating system is that a substantial fraction of the participants in the system behave altruistically by supplying honest evaluations. Still, we predict that a credibility-weighted tool would provide the user with a means to filter out ratings of dubious nature. The use of a credibility-weighted tool is also advantageous for the subcontractors who are rated in the system. Some of them may object to being rated out of fear of receiving "unjust" ratings from ill intending general contractors. If the system weighs the ratings by rater credibility the subcontractors can feel certain that the effects of any dishonest ratings will be limited. A general contractor who provide inaccurate ratings are less likely to be perceived as credible, and as a result their ratings will have less impact on the overall ratings. As a result, a credibility-weighted rated tool could be an important building block in a framework to create trust in AEC e-commerce. By rating the rater the user of AEC e-market places can come one step closer to trusting new business partners.

REFERENCES

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

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

Birnhaum, M. H., and Stegner, S. E. (1979). "Source Credibility in Social Judgment: Bias, Expertise and the Judge's point of view." *Journal of Personality and Social Psychology*, 37(1), 48-74.

Chen, L.-H., and Chiou, T.-W. (1999). "A fuzzy credit-rating approach for commercial loans: a Taiwan Case." *Omega, International Journal of Management*, 27, 407-419.

Cohen, L., and Holliday, M. (1982). *Statistics for social scientists*, Harper & Row Ltd, London.

Fisher, C. D., Ilgen, D. R., and Hover, W. D. (1979). "Source Credibility, information favorability, and job offer acceptance." *Academy of Management Journal*, 22(1), 94-103.

Harmon, R. R., and Coney, K. A. (1982). "The persuasive effects of source credibility in buy and lease situations." *Journal of Marketing Research*, 19, 255-260.

Hovland, C. I., Janis, I. L., and Kelley, H. H. (1953). *Communication and Persuasion*, Yale University Press, New Haven.

Hovland, C. I., and Weiss, W. (1951). "The Influence of Source Credibility on Communication Effectiveness." *Public Opinion Quarterly*.

Howard, R. A. a. J. E. M. (1984). "Readings on the Principles and Applications of Decision Analysis." , Strategic Decisions Group, Menlo Park.

McCroskey, J. C. (1966). "Scales for the measurement of ethos." *Speech Monographs*, 33(1), 65-72.

Rice, J. A. (1995). *Mathematical Statistics and Data Analysis*, Wadsworth Publishing Company, Belmont.

Romaniuk, S. G., and Hall, L. O. (1992). "Decision making on creditworthiness using a fuzzy connectionist model." *Fuzzy Sets and Systems*, 15-22(48).

Saaty, T. L. (1980). *The Analytic Hierarchy Process, Planning, Priority, and Resource Allocation*, McGraw-Hill, New York.

Stone, E. F., and Stone, D. L. (1984). "The effects of multiple sources of performance feedback and feedback favorability on self perceived task competence and perceived feedback accuracy." *Journal of Management*, 10, 371-378.

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

Uebersax, J. S. (2001). "Raw Agreement Indices." .

Zacharia, G., Moukas, A., and Maes, P. "Collaborative Reputation Mechanisms in Electroni Marketplaces." *Thirty-second Annual Hawaii International Conference on System Sciences (HICSS-32)*, Wailea, Hawaii.

Zimmerman, H.-J., and Zysno, P. (1983). "Decision Evaluations by hierarchical aggregation of information." *Fuzzy Sets and Systems*, 10, 243-260.