Rating System Design: Transforming Individual Preferences to Rating Scores

Anna E. Bargagliotti
Department of Mathematical Sciences
University of Memphis
abargag@yahoo.com

Lingfang (Ivy) Li
Department of Economics
University of Louisville
ivy.li@louisville.edu

June 27, 2008

Abstract

Rating systems measuring quality of products and service (i.e. the state of the world) use ranking methods in order to solve the asymmetric information problem in markets. Different metrics and data aggregation procedures may translate the same underlying popular opinion to different conclusions about the true state of the world. This paper characterizes the differences in metric systems used in the different rating systems by defining regions of an individual's perceived states in the interval [1,N] that are consistent regardless of what metric is being used. It is shown that the only scaled metric (1-N) that reports people's opinion equivalently in the a binary metric (-1, 0, 1) is one where N is odd and N-1 is not divisible by 4. Differences in the data aggregation systems are characterized and simple tools are illustrated that determine whether different rating systems are consistent. In addition, this paper also provides answers to questions regarding when and how often the systems differ. [1] [2]

[1]JEL: D82, D70. Keywords: rating, ranking, preference, asymmetric information.

[2]The author names are in alphabetical order, and both authors equally contributed to this paper. Thank you to Donald Saari for his comments on earlier drafts.

1 Introduction

Rating systems are widely used in business, political systems, and in our daily lives as methods of containing and communicating information that are crucial to making decisions. For instance, investors make investment decisions by looking at ratings of financial products, online shoppers compare products by looking at the seller and product ratings, doctors make decisions based on patient's ratings of their well-being, patients search for doctors by looking at their score on MD rating sites, students refer to US News University rankings to decide which school to attend, and universities use rating systems to monitor professors' performance in teaching. Rating systems provide a way to summarize public opinion in an organized manner. Each rating system is defined by a rating scale and a rule that aggregates individual's ratings into a single overall score. Several forms of rating systems exist. For example, eBay's reputation system asks users to rate a buyer or seller in a binary rating system (positive, neutral, and positive). On the other hand, Amazon.com asks raters to use a scaled rating system of one through five where one is considered worst and five is best. Hotels.com provides users both a one through five scaled rating system and a binary option. The binary metric is described by a "recommend" option or a "not recommend" option. Questions arise regarding which type of metric best represents public opinion as well as whether the binary and scaled metric represent information in the same manner.

Assuming that individuals express their opinion truthfully,[3] different rating systems may summarize opinions differently. For example, in the case of Hotels.com, a person may be confronted with leaving a rating in a one through five scale as well as a binary system with options described as "recommend" or "not recommend." A person may assign a rating of three in the scaled metric, but in the binary metric, choose "not recommend." These are not equivalent statements. In the scaled metric, the number three represents a neutral rating, while "not recommend," is negative. Similar issues occur in rating stocks, sellers, or products. It is not uncommon to observe inconsistent scenarios. Consider two groups of ten financial analysts rating a stock. On a scale of one to five, five analysts in the first group rate the stock as a 4 and the other five rate it as a 2 in the scaled

[3]This paper assumes no strategic rating takes place in the systems. This means that every individual leaving a rating rates according to their actual beliefs

system. When asked to rate the stock in the binary system (buy, hold, and sell), the aggregate group decision is "buy." Suppose in the second group all ten people rate the stock as a 3 in the scaled system but in the binary system, the group decides to "hold." In both the scaled cases, the overall opinion of the ten people seems to be that the stock is neutral. However, in the binary system, the results are quite different. Should the two cases and two metrics produce equivalent results? Why do these differences occur? It is hypothesized that these types of inconsistencies are due to human error. For example, a person may not understand the meaning of the scale values. A person may not recognize that the rating three is actually a neutral rating and thus would not recognize that their rating for "recommend" and "buy" or "not recommend" and "sell" would inflate or deflate their opinion. However, the person could understand the scales perfectly but still rate inconsistently in the two systems due to the restrictions of their metrics.

This paper investigates whether certain rating schemes have the ability to represent public opinion more accurately than others. In particular, the scaled 1-N and binary system (-1,0,1) are compared and their differences are characterized. This is done by first understanding how the two metrics represent an individual's opinion. Implications of the different representations, as well as when and how often the differences occur at the individual level, are discussed. In addition, the best possible scaled metric is identified. Secondly, each rating system is partially defined by an aggregation rule. The effect of the different aggregation rules in the given metric is examined. The differences in the aggregation systems are characterized and simple tools are illustrated to determine whether the binary and scaled rating systems are reporting the same information. In addition, the frequency of the inconsistencies are discussed at the aggregation level in terms of a probability distribution. Questions regarding when and how often the systems differ at the aggregation level are answered.

2 Related Literature

Research has focused on obtaining best methods for summarizing information and public opinion. Ranking methods and rating methods have been proposed as ways to accomplish this. In general, a ranking system is a method that asks participants to rank-order alternatives while a rating system

asks participants to score an alternative on an arbitrary scale. In both cases, the alternatives could represent the object of opinion or the opinion itself. Droba (1931) provides a review on methods used for measuring public opinion in social science that include both methods: ranking and rating. The author points out that in the case where the opinion itself is the object being summarized, ranking methods could be a special case of rating systems. For example, a person may be asked to organize alternatives into three different groups which represent the worst, middle, and best alternatives. The categories can be represented by arbitrary numbers such as, worst may be 0, middle as 1, and best as 2. In this case, the ranking method is asking the participant to rank according to a rating. This types of ranking methods is actually providing ratings. Stevens (1946) discusses the types of mathematical group structures and statistical operations a rating scale could have. He classifies the scales of measurement into nominal, ordinal, interval, and ratio measurement categories. Alwin and Krosnick (1985) provide a comparison study of rating and rankings in terms of measuring values in surveys. The ranking approach orders a set of competing alternatives while the rating approach rates alternatives on a scale of importance. The authors provide both pros and cons for each approach. One drawback of using the ranking method is that the task of rank-ordering alternatives is very difficult for respondents to complete. Obtaining a total ordering on a large set of alternatives, may be a daunting task for a participant while rating each alternative individually may be easier. On the other hand, rating methods do not differentiate well among people's opinion. It has been observed that respondents tend to rate everything high. This issue makes the system unable to differentiate well among the various values. It is hypothesized that this phenomena occurs because people may not understand the meaning of the scale they are presented with. Although rating and ranking methods are both used to capture the population's opinion, this general literature makes a distinction between the two. According to the definitions provided in the literature, this paper focuses on the structure of rating systems.

Although the mathematical structure of rating systems has not been explicitly studied in Economics, the social choice implications of voting systems have been intensely explored by economists, decision scientists, and political scientists. The problem of aggregating preferences is one of the most vexing and difficult in these fields. Issues of existence of aggregation functions go back at least as far as Arrow's Impossibility Theorem. Since then, numerous papers have discussed voting

and social choice rules used for aggregating preferences.[4] To fully understand the source of the aggregation problems, recent research papers in social choice have focused on examining the mathematical structure of social choice rules that are used to aggregate individual's preferences. Saari (1999, 2001a, 2001b) provide mathematical tools to revisit the paradox in voting and impossibility theorems in social choice. He provides a new interpretation of Arrow's impossibility results. For instance, Arrow's Theorem has a benign interpretation. Li and Saari (2008) provide the first direct mathematical proof of Sen's seminal result in social choice. The proof underscores a significantly different interpretation of the driving reasons of Sen's theorem. Rather than conflicts among voter rights, the proof illustrates that the liberalism assumption negates the requirement that voters have transitive preferences. Suggestions on how to sidestep these difficulties are built upon the new findings.

Rating systems are also an important part of market reputation systems. A reputation system is an important institution that helps sustain trust in the market. Especially in online markets, where Internet traders are anonymous and geographically dispersed, rating systems are used in order to monitor and establish stability. Research in Economics and Management Information Systems focuses on exploring the fairness of reputation systems in online systems. For example, eBay's current reputation system has problems of biased rating towards positive feedback. [5] Suggestions of various ways to repair these problems have been made.[6] Dellarocas et al. (2006) provide a comprehensive review on topics related to reputation systems. They give suggestions on improving reputation metrics and the aggregation rule of ratings in reputation system. For example, using non-negative feedback scales like 0,1 or 0,1,2 instead of negative scales like -1,0,1 may prevent loss of participation in the market. Another improvement may be using the sum of ratings provided by single users in the most recent N transaction instead of over the whole history of transactions. Dellarocas (2003b) finds that introducing a more complex rating system, such as a scaled system, cannot improve the reputation mechanism's efficiency in terms of high level of cooperation between traders.

[4]See Arrow (1963), Moulin (1988), Sen (1970, 1986), and Saari (2001a, 2001b) for references and surveys on the topics.

[5]See Dellarocas and Woods (2006) and Klein et al. (2006).

[6]See Miller et al (2005), Jurca and Faltings (2004), Li(2007)

Particular applications of rating system research exists in Marketing and Psychology. Several observations in these disciplines illustrate that people rate inconsistently when rating in different metrics. Much of this literature argues that the inconsistence between different scaled systems are due to psychological reasons or other human errors such as misunderstanding the descriptions of the scales. In Marketing research, choosing a rating scale is an important part of designing a study to build marketing effectiveness.[7] Researchers in Marketing find that people tend to assign high ratings to all items in the set. This may be due to the fact that people do not differentiate greatly among the various values.[8] This type of phenomena makes the rating data low quality since it does not provide much information about people's opinions.

Psychology researchers examine variables ranging from rating scale formats to the cognitive limitations of raters. These have been identified as potential explanations for the apparently low quality of rating data. For example, Saal, Downey, Lahey (1980) provide a survey of research on examining rater's error in terms of individual's behavior and methods that quantify rater's errors. Landy Farr (1980) reviews literature on performance rating. They argue that the cognitive characteristics of raters seem to explain the bias toward high ratings in rating systems. Schwarz et al. (1991) use the example that a scale from -3 to 3 creates different results than a scale from 1 to 7. They find that the numeric values provided as a part of rating scale influence respondents' interpretation of the end point labels. Cleveland and Murphy (1992) and Murphy et al. (2004) suggest that the reason for raters to give ratings that appear psychometrically suspect is due to their different goals when completing performance appraisals. For example, raters may want to maintain harmony within the workgroup or possibly motivate subordinates to perform better in the future. A rater who gives high ratings might not be making a judgment error but rather might be making a decision that it is better to give all subordinates high ratings than to give low ratings to poor-performing subordinates. High ratings might lead to better pay for subordinates, more harmony in the workgroup, and better supervisorsubordinate relations. On the other hand, more accurate ratings may lead to resentment, low motivation, and friction in the group. They introduce these ideas to explain reasons for low quality rating results.

[7] see McDaniel and Gates 2004, Lehmann and Hulbert 1972.

[8] See McCarty and Shrum 2000, Greenleaf, Bickart, and Yorkston 1999.

Reviewing the current literature, it can be seen that the existing explanations of inconsistencies in rating systems focus mostly on possible human errors. This paper shows that even without human errors, the inconsistency among rating systems can occur due to the mathematical construction of the rating systems. The focus of the current paper is to characterize the differences in the systems in terms of the metrics. While previous contributions explain the "meaning" of the scales and how people interpret them, this paper strives to understand the structural differences among the scales.

3 Individual Rating Using Different Metrics

A rating system is composed of two separate items: a metric and an aggregation rule. Individuals rate in the metric and then their scores are aggregated according to the specified rule. The metric is typically a discrete set of integers contained in an interval. For example, the scale metric 1,2,3,...,10 can be expressed as the set of integers in the interval [1,10]. An aggregation rule combines people's opinions into one overall score. Typically, the rule takes the form of an averaging function. Mathematically, a rating system can be defined in the following manner.

Definition 1 *Let* $int[m,n]^k$ *be the k-product of the set of integer values in the interval [m,n]. A rating system is a set* { *int[m,n], R* } *where int[n,p] is the set of integers between m and n containing m and n. R is a function from* $int[m,n]^k$ *to the interval [m,n] and k is the number of people rating.*

$$R\text{: } int[m,n]^k \rightarrow [m,n]$$

The binary rating system can therefore be specified as { int[-1,1], R_b } where int[-1,1] = {-1, 0, 1 } and R_b: $int[a,b]^k \rightarrow [a,b]$. The scaled system can be written as { int[1,N], R_s } where int[1,N] = { 1, 2, 3, ..., N } and R_s: $int[1,N]^k \rightarrow [1,N]$.

The binary and scaled rating systems use different metrics. How do these different metrics affect the manner in which people vote? Can both systems represent an individual's opinion? If so, do they represent the opinion in the same manner?

Whenever a trade takes place, there exists a true state, TS, of the trade. TS can be thought of as the true quality of the trade. For example, if students are asked to rate their professors, there exists a true quality of the lessons removed from the student's opinion. In an on-line market, a seller will rate a buyer according to their satisfaction of the transaction, however, a true quality of the transaction exists outside of the buyers opinion. Because people typically people think in a positive continuous interval, the true state TS is modeled as an point in the interval [1,N].

Each individual perceives TS differently. An individual forms a perceived state, PS, that represents their opinion of TS. Similarly to TS, PS lies in [1,N]. For the purpose of this paper, it is assumed that people report honestly, according to their actual perception of the true state. This means that given a metric an individual rates at the closest available rating to their PS.

Example 1 *Suppose two people are rating a transaction in the interval [1,5] given the rating system* $\{$ *int[1,5],* $R_s\}$*. Person A perceives the transaction as 2.1 and person B perceives it at 1.9. Since the available rating systems requires the individuals to rate 1, 2, 3, 4, or 5 both people vote 2. Both 2.1 and 1.9 perceived states are closest to 2.*

How would the individuals translate their opinions to rate in the binary system? The metric would change from int[1,5] to int[-1,1] in the binary system. Rating in the binary system requires an individual to translate their opinion from [1,N] to [-1,1]. Once this step is complete then the individuals will again rate at the closest available option to their PS in the binary system. The perceived state in the interval [1,N] can be translated to a point in the interval [-1,1] using a linear transformation. The following definition provides an equation to translate values between the binary and scaled systems.

Definition 2 *Given systems* $\{$ *int[a,b],* R_1 $\}$ *and* $\{$ *int[c,d],* R_2 $\}$*. Any point k in [a,b] can be transformed to a point j in [c,d] in the following manner:*

$$k=\frac{da-cb}{d-c}+\frac{b-a}{d-c}\ j \qquad (1)$$

This means that a perceived state PS_1 in system 1 can be transformed into a perceived state PS_2 in system 2 by:

$$PS_1=\frac{da-cb}{d-c}+\frac{b-a}{d-c}PS_2 \qquad (2)$$

A rating r_1 in system 1 can be transformed to a rating r_2 in system 2 in the following manner:

$$r_1=\frac{da-cb}{d-c}+\frac{b-a}{d-c}r_2 \qquad (3)$$

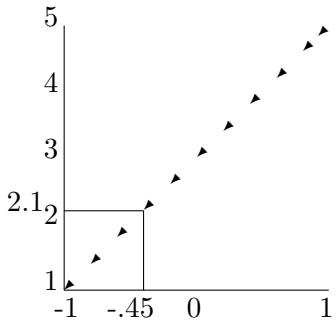
where $r_1=\{k : k\in$ [1,N] and min$|PS_1\text{-}k|\}$ and $r_2=\{j : j\in$ [1,N] and min$|PS_2\text{-}j|\}$. PS_1 and PS_2

are equivalent perceived states in systems 1 and 2.

Therefore a perceived state, PS, in the scaled system [1,N] is equivalent to perceived state in the binary system [-1,1] if $PS_{[1,N]}=\frac{1+N}{2}+\frac{N-1}{2}PS_{[-1,1]}$.

Example 2 *Person A and person B have the same perceived states, 2.1 and 1.9 respectively as example above. Then 2.1=*$\frac{1+5}{2}+\frac{(5-1)}{2}PS_{[-1,1]}$*=3+2(*$PS_{[-1,1]}$*) and thus person A has perceived state in [-1,1] equal to -.45. In the same manner, person B has perceived state in [-1,1] equal to -.55. Person A and B may rate -1, 0, or 1 in this metric. Since -.45 is closest to 0, A rates 0. Since -.55 is closest to -1, B rates -1.*

Figure 1: Transforming scores between metrics



Examples 2 illustrates how scored can be translated from the scaled to the binary system. Interestingly, examples 1 and 2 point out how the two systems represent their opinions differently. In example 1, in the scaled metric, person A and B rate the same, however, example 2 illustrated that the translated perceived scores in the binary metric cause A and B to rate differently. The only thing being varied in the examples is the metric. The people's perceptions and the true state remained the same but their opinions are represented differently. Are there specific perceived states that are problematic? When are the two systems equivalent? When do they lead to different results?

Definition 3 *A scaled system and a binary system are called equivalent if and only if any two*

raters who rate the same in scaled system implies they also rate the same in binary system.

In other words, we consider the scaled and the binary system equivalent if they represent people's opinion in the same way. This means that if the scaled system represents two raters opinions as the same then the equivalent binary system must also represent the raters opinions as equal.

According to definition 3, in order for the binary and the scaled system to be equivalent, two individuals that have the same rating in the scaled rating systems must have the same rating in the binary system as well. According to definition 2, the perceived states and ratings can be translated between the two systems using a linear function. The examples above illustrate that the binary and scaled metrics are not equivalent.

3.1 Characterization of Differences between Binary and Scaled Metric at the Individual Rater Level

Are there conditions that might force the binary and scaled metrics to be equivalent? Is there something particular about the perceived states in the examples above that lead to the inconsistent results? The following theorems address these inconsistencies by specifying regions of the interval [1, N] on which the metrics produce equivalent results and regions that lead to inconsistent outcomes.

Theorem 1 *For any odd number N where $N-1$ is divisible by 4, the scaled system and the binary system are equivalent if and only if the two individual's perceived score are within* $[1,N]\backslash[\frac{1+\frac{1+N}{2}}{2} - \frac{1}{2}, \frac{1+\frac{1+N}{2}}{2} + \frac{1}{2}] \bigcup [\frac{\frac{1+N}{2}+N}{2} - \frac{1}{2}, \frac{\frac{1+N}{2}+N}{2} + \frac{1}{2}]$. *In particular, the systems will not be equivalent if one rater has perceived score to the left of* $\frac{1+\frac{1+N}{2}}{2}$ *and the other to the right in the given interval. Or, the systems will not be equivalent if one rater has perceived score to the left of* $\frac{\frac{1+N}{2}+N}{2}$ *while the other has a perceived score to the right in the given interval.*

Theorem 2 *For any odd number N where $N-1$ is not divisible by 4, the scaled system and the binary system are equivalent everywhere.*

Theorem 3 *For any even number N where N is divisible by 4, the scaled system and the binary system are equivalent if and only if the two individual's perceived states are within* $[1, N]\backslash[\frac{1+\frac{1+N}{2}}{2} - \frac{1}{4}, \frac{1+\frac{1+N}{2}}{2} + \frac{3}{4}] \bigcup [\frac{\frac{1+N}{2}+N}{2} - \frac{3}{4}, \frac{\frac{1+N}{2}+N}{2} + \frac{1}{4}]$. *In particular, the systems will not be equivalent if one rater has perceived score to the left of* $\frac{1+\frac{1+N}{2}}{2}$ *and the other to the right in the given interval. Or, the systems will not be equivalent if one rater has perceived state to the left of* $\frac{\frac{1+N}{2}+N}{2}$ *while the other has a perceived score to the right in the given interval. If N is greater than 2 and not divisible by 4, the scaled system and the binary system are equivalent if and only if the two individual's perceived states are within* $[1, N]\backslash[\frac{1+\frac{1+N}{2}}{2} - \frac{3}{4}, \frac{1+\frac{1+N}{2}}{2} + \frac{1}{4}] \bigcup [\frac{\frac{1+N}{2}+N}{2} - \frac{1}{4}, \frac{\frac{1+N}{2}+N}{2} + \frac{3}{4}]$. *If $N = 2$, the the two systems are equivalent if and only if the two individual's perceived states are in* $[1, 1.25]$ *and* $[1.75, 2]$ *as in the 1-2 scaled system.*

Figure 2: N is odd and N-1 divisible by 4 characterization

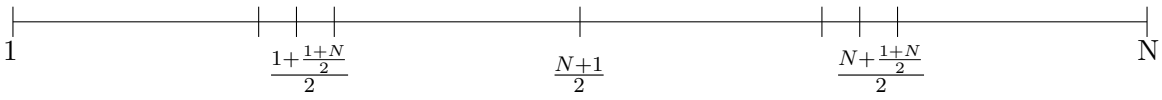


Figure 3: N is even characterization and divisible by 4

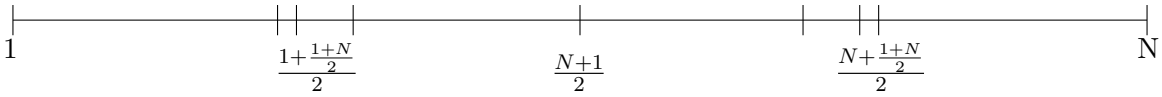
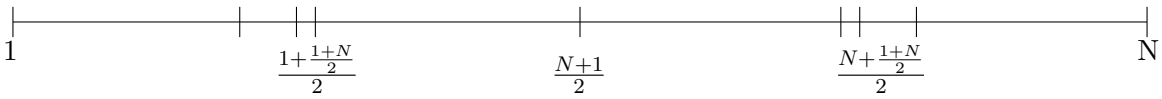


Figure 4: N is even characterization and not divisible by 4



The figures and theorems illustrate the regions where the systems represent people's opinions in different manners. One can see that in the case where N is odd and N-1 is divisible by 4, for example, if the two raters have perceived scores on either side of $\frac{1+\frac{1+N}{2}}{2}$ but within $[\frac{1+\frac{1+N}{2}}{2} - \frac{1}{2}, \frac{1+\frac{1+N}{2}}{2} + \frac{1}{2}]$

, then in the scaled system both raters rate $\frac{1+\frac{1+N}{2}}{2}$ but in the binary system, they rate -1 and 0. The number $\frac{1+\frac{1+N}{2}}{2}$ essentially "splits" the raters in the binary system. Therefore, in order to ensure the equivalency of the systems, both raters must have perceived scores not in these critical regions. This characterization points out how the metrics interpret information differently and cause problems to occur.

3.2 Implications of Differences between Binary and Scaled Metric at the Individual Rater Level

The binary and the scaled metric interpret ratings in a different manner except for in the case where N is odd and is not divisible by 4. What are the implications of this result? What are the implications of the differences between the systems for the cases other than N is odd and divisible by 4? Do the systems tend to penalize or inflate each rating in these cases? As seen in example 2 above, the binary and the scaled system are not equivalent at the individual level for all N. Persons A and B rate the same in the scaled system but do not in the binary system. In fact, in the example, person A's rating of 0 in the binary system actually is higher than his/her score in the scaled. Person A rated 2 in the scaled system, however, a rating of 0 in the binary system is a neutral score not a negative score. This means that a 0 rating is actually the same as a rating of 3 in the int[1,5] scaled system. This means that the rating of 2 is lower than it's counterpart of 0. With perceived state equal to 2.1, the binary system actually inflates the score to 0. To the contrary, person B's perceived state is penalized in the binary system. Person B rates -1 in the binary system which is not equivalent to 2 in the scaled. Having a perceived state equal to 1.9 actually causes a distortion penalizing the truth. The following theorem characterizes the regions of the interval [1, N] that the binary system distorts.

Theorem 4 *For all N, let PS_s equal a rater's perceived state in the scaled system and r_s be it's equivalent rating given by r_s=$\{k : k\in$ [1,N] and $min|PS_s\text{-}k|\}$. Let r'_{sb} equal the translation of r_s into the binary system. Define PS_b be the translation of PS_s in the binary system and r_b be it's rating given by r_b=$\{j : j\in$ [-1,1] and $min|PS_b\text{-}k|\}$. The penalty or benefit associated with every*

rating is given by r_b-r'_{sb} if using the binary system instead of the scaled system.

Notice if N is odd and divisible by 4 and PS_s=1, $\frac{1+N}{2}$, or N, then r_b-r'_{sb}=0. If N is even and PS_s=1 or N, then r_b-r'_{sb}=0. This theorem illustrates that for all but a few values, the binary system represents people's opinions in a different manner than the scaled system. These limitations of the metric make it so that a person's perceived state is inflated or deflated in the binary system. In order to compute that amount of distortion created by the metric, the difference in what the rating should be, i.e. the translated ranking r'_{sb}, and the actual binary rating r_b is computed. The following example illustrates the computation.

Example 3 *Suppose N=5. Let $PS_s \in [1.5, 2]$ be given. Then the individual will rate r_s= 2 in the int[1,5] metric. The 2 rating in the scaled system is equivalent to the position at r'_{sb}=-0.5 in the binary system. However, in the binary system, an individual with $PS \in [1.5, 2]$ will rate r_b=-1. Therefore, r_b-r'_{sb}= -0.5 penalty in the binary system.*

If $PS \in [2, 2.5]$, the individual will rate as a r_s= 2 in the scaled system. As before, the rating 2 is equivalent to the position r'_{sb}=-0.5 in the binary system. However, for $PS \in [2, 2.5]$, the individual will rate r_b=0. Therefore, r_b-r'_{sb}= 0.5 benefit in the binary system.

Theorem 5 *The best type of scaled system is one that has an odd N and where N-1 is not divisible by 4.*

Due to Theorem 2, this is the only scenario in which the two systems are equivalent on all regions at the individual rater level. This means that this is the only case where the binary system does not distort the scaled system rating. In this case, the two systems represent a person's individual opinion in the same manner. This result underscores the importance of scale selection in applications. For example, in the case of Hotels.com or stock rating described in the example, raters are asked to rate both on a binary metric and a scaled metric. Because the website asks raters to rate on a scale of one through five, then there is no way that raters can maintain equivalent results between the two metrics. This result explains why these inconsistencies occur at the individual level in applications.

4 Aggregation Procedures Using Different Metrics

Once individuals have cast their ratings, the information is aggregated to form an overall score for the event. For example, in a university setting at the end of a semester a faculty member is evaluated by his/her students. Each student answers a series of questions, typically on a scale, and then the answers are aggregated to an overall class rating of the professor. In on-line markets, buyers and sellers have their transaction history summed up into an overall score. Many different aggregation rules can be found in current market societies. Typically, R is defined as an average or weighted average function.

Example 4 *Suppose there are three students rating professor Smith. Each student is asked to rate Dr. Smith on a scale of 1-4. The following table illustrates the student's perceived states on the interval [1,4] and their respective ratings in the scaled metric.*

Students	Perceived State in [1,4]	Scaled Metric Vote
Student 1	*1.1*	*1*
Student 2	*2.7*	*3*
Student 3	*3.1*	*3*

Then R *:int[1,4]*$^3 \rightarrow [1,4]$ *aggregates scaled scores 1, 3, and 3 by averaging:* $\frac{1+3+3}{3}$*=2.5*

In the previous sections, the differences between the binary and scaled systems were explored and characterized. In order to compare the metrics further, this section focuses on exploring how the average function R aggregates in the two metrics. This aggregation level comparison is done in the best case scenario: the rule is defined as the average function in both metrics. It may be the case that different rules are actually used in practice, however, the simplest comparison is explored in this paper. Inconsistencies and differences between rating systems also occur at the aggregation stage for the case where the average function is used in both metrics.

4.1 Formulation of scores

In the binary system, let b_1, b_2, and b_3 represent the number of -1, 0, 1 ratings respectively. In the scaled system, let s_1, s_2,..., s_N represent the number of 1, 2, ..., N ratings respectively. The aggregation rule R in both systems then takes the k individual ratings and computes b_1, b_2, b_3 or s_1, s_2,..., s_N respectively for each systems. Once this computation is complete, the average function R_b aggregates the scores the binary system as: $\frac{b_3-b_1}{b_1+b_2+b_3}$ and R_s aggregates in the scaled system as: $\frac{\sum_{i=1}^{N} is_i}{\sum_{i=1}^{N} s_i}$. Thus, R_b is defined as a function from $int[-1,1]^k \rightarrow [-1,1]$ and R_s is a function from $int[1,N]^k \rightarrow [1,N]$.

4.2 Characterizing the Differences in Aggregation Procedures

Both R_b and R_s are defined to be the average function on their respective metrics. Comparing the two aggregation rules amounts to understanding what the values of b_1, b_2, b_3 and s_1, s_2,..., s_N must be in order for the two rules to output equivalent scores for the event.

Definition 4 *Let R_x: $int[a,b]^k \rightarrow [a,b]$ and R_y: $int[c,d]^k \rightarrow [c,d]$. R_x and R_y are consistent if and only if the aggregated score in one metric is equal to a linear transformation of the aggregated score in the other metric. i.e. R_x and R_y satisfy the following equation:*

$$R_y = \frac{bc-da}{b-a} + \frac{d-c}{b-a} R_x. \qquad (4)$$

Because different aggregation procedures have different domains and ranges, they are only considered equivalent if they output equivalent scores. Determining if they output equivalent scores requires a similar linear transformation to the one given in definition 2 between the outputs of

the aggregation functions. In the case of the binary and scaled systems, they are consistent provided the aggregated binary score is equal to the linear translation of the aggregated scaled score. This means rules R_b and R_s where R_b:$int[-1,1]^k \rightarrow [-1,1]$ and R_s:$int[1,N]^k \rightarrow [1,N]$ satisfy the following equation:

$$\frac{\sum_{i=1}^{N} is_i}{\sum_{i=1}^{N} s_i} = \frac{1+N}{2} + \frac{N-1}{2}\left(\frac{b_3 - b_1}{b_1 + b_2 + b_3}\right). \tag{5}$$

Here are two examples that explore the consistency of R_b and R_s.

Example 5 *Suppose two people rate the same event in both a 1-5 scaled system and the binary system -1,0,1. The following table lists their perceived states and their respective score in each metric.*

Raters	Perceived State in [1,5]	Scaled Metric Vote	Binary Metric Vote
Rater 1	*1.1*	*1*	*-1*
Rater 2	*2.7*	*3*	*0*

The aggregation rule in the binary system, computes an overall score of $R_b = \frac{-1+0}{2} = \frac{-1}{2}$ *and the rule in the scaled system computes* $R_s = \frac{1(1)+3(1)}{2} = 2$. *These aggregation results are consistent by Eq.(5).*

Are R_b and R_s consistent for all possible perceived states? The next example illustrates that the average function in the two metrics are not consistent given certain rater perceived states.

Example 6 *Again, the following table lists the perceived states of the two raters. They differ from the previous example only in the first rater's perceived state.*

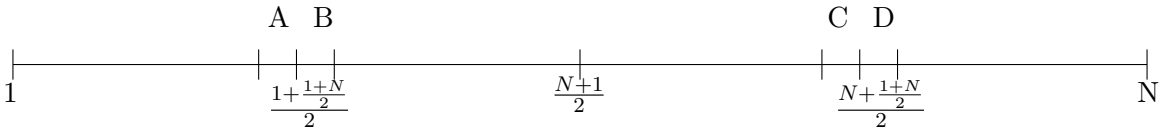
Raters	Perceived State in [1,5]	Scaled Metric Vote	Binary Metric Vote
Rater 1	*1.6*	*2*	*-1*
Rater 2	*2.7*	*3*	*0*

The binary rule tallies $R_b = \frac{-1+0}{2} = \frac{-1}{2}$ *and the scaled rule tallies* $R_s = \frac{2(1)+3(1)}{2} = 2.5$. *These aggregation results are not consistent by to Eq. (5).*

These examples show that consistency between R_b and R_s is dependent on the values of b_1, b_2, b_3 and s_1, s_2,..., s_N. In turn, these values are dependent on the number of rater's perceived states in certain regions of the ranges of the aggregation functions [-1,1] and [1,N]. How do the rater's perceived states influence the aggregation outcomes? Can perceived state regions characterize when the aggregation rules are consistent?

Definition 5 *Given the interval [1,N] where N is odd, define region* $A=(\frac{1+(\frac{N+1}{2})}{2}-\frac{1}{2}, \frac{1+(\frac{N+1}{2})}{2})$, *region* $B=(\frac{1+(\frac{N+1}{2})}{2}, \frac{1+(\frac{N+1}{2})}{2}+\frac{1}{2})$, *region* $C=(\frac{N+(\frac{N+1}{2})}{2}-\frac{1}{2}, \frac{N+(\frac{N+1}{2})}{2})$, *and region* $D=(\frac{N+(\frac{N+1}{2})}{2}, \frac{N+(\frac{N+1}{2})}{2}+\frac{1}{2})$. *Let* $|X|$ *denote the number of voter's perceived states in region X.*

Figure 5: N is odd regions



The following theorems characterize the differences in aggregation procedures by stating conditions the on rater's perceived states in order to ensure consistency between R_b and R_s.

Theorem 6 *Let N be odd and regions A, B, C, and D be defined as in definition 5. If* $|A|=|B|$ *and* $|C|=|D|$ *then* R_b *and* R_s *will be consistent.*

If N-1 is divisible by 4, then perceived states in regions A and B in the scaled system are rated as $\frac{1+\frac{N+1}{2}}{2}$. In the binary system, states in region A are rated as -1 and states in region B are rated

as 0. Perceived states in regions C and D in the scaled systems are rated as $\frac{N+\frac{N+1}{2}}{2}$, while in the binary system region C is rated as 0 and region D is rated as 1. If A and B have the same number of perceived states, there there are |A|+|B| voters rating at $\frac{N+\frac{N+1}{2}}{2}$ which is halfway between 1 and the midpoint $\frac{1+N}{2}$. This means that the aggregation function R_s will have a value of exactly $\frac{N+\frac{N+1}{2}}{2}$. Equating the number of perceived states in region A and B ensures that the average score in the binary system will be at $\frac{-1}{2}$ which is equivalent to the average score of the halfway point of $[1,\frac{1+N}{2}]$ in the scaled system. Similarly, the situation holds for regions C and D. If instead N-1 is not divisible by 4, then perceived states in region A will rate at $\frac{1+(\frac{N+1}{2})}{2}$-$\frac{1}{2}$ while perceived states in B will rate at $\frac{1+(\frac{N+1}{2})}{2}$+$\frac{1}{2}$. Similarly to the divisible by 4 case, having |A|=|B| ensures that R_s will have a value of exactly $\frac{1+\frac{N+1}{2}}{2}$. This is consistent with the binary aggregation that scores $\frac{-1}{2}$.

Theorem 7 *Let N be odd and regions A, B, C, and D be defined as in definition 5. If |A|=|D| and |B|=|C| then R_b and R_s will be consistent.*

Theorem 7 exemplifies how a "reverse" type symmetry in the perceived states ensures consistency between R_b and R_s. For the situation described in the theorem, the scaled system aggregation function R_s will equal the midpoint $\frac{1+N}{2}$. This symmetric balance is carried over to the binary system since the raters in region B and C rate 0 while the voters in region A rate -1 and those in region D rate 1. If there are an equal number of voters who have perceived states at -1 and 1 then the average aggregated score is 0. Scoring 0 in the binary system is exactly equivalent to scoring the midpoint $\frac{1+N}{2}$ in the scaled system.

Definition 6 *Given the interval [1,N] where N is even and divisible by 4, define region A=($\frac{1+\frac{1+N}{2}}{2}-\frac{1}{4}$, $\frac{1+(\frac{N+1}{2})}{2}$), region B=($\frac{1+(\frac{N+1}{2})}{2}$, $\frac{1+\frac{1+N}{2}}{2}+\frac{3}{4}$), region C=($\frac{\frac{1+N}{2}+N}{2}-\frac{3}{4}$, $\frac{N+(\frac{N+1}{2})}{2}$), and region D=($\frac{N+(\frac{N+1}{2})}{2}$, $\frac{N+(\frac{N+1}{2})}{2}+\frac{1}{4}$). If N is greater than 2 and not divisible by 4, define region A=($\frac{1+\frac{1+N}{2}}{2}-\frac{3}{4}$, $\frac{1+(\frac{N+1}{2})}{2}$), region B=($\frac{1+(\frac{N+1}{2})}{2}$, $\frac{1+\frac{1+N}{2}}{2}+\frac{1}{4}$), region C=($\frac{\frac{1+N}{2}+N}{2}-\frac{1}{4}$, $\frac{N+(\frac{N+1}{2})}{2}$), and region D=($\frac{N+(\frac{N+1}{2})}{2}$, $\frac{N+(\frac{N+1}{2})}{2}+\frac{3}{4}$). Let |X| denote the number of voter's perceived states in region X.*

The following theorems characterize the differences in aggregation procedures by stating conditions the on rater's perceived states in order to ensure consistency between R_b and R_s when N is an even

number.

Theorem 8 *Let N be even and regions A, B, C, and D be defined as in definition 6. If* $|A|=|D|$ *and* $|B|=|C|$ *then* R_b *and* R_s *will be consistent.*

Theorem 8 follows the same argument as in Theorem 7. It exemplifies how a "reverse" type symmetry in the perceived states ensures consistency between R_b and R_s. For the situation described in the theorem, the scaled system aggregation function R_s will equal the midpoint $\frac{1+N}{2}$. This symmetric balance is carried over to the binary system since the raters in region B and C rate 0 while the voters in region A rate -1 and those in region D rate 1. If there are an equal number of voters who have perceived states at -1 and 1, then the average aggregated score is 0. Scoring 0 in the binary system is exactly equivalent to scoring the midpoint $\frac{1+N}{2}$ in the scaled system.

Theorem 9 *If N is odd, define regions A, B, C, and D as in definition 5. If N is even, then define regions A, B, C, and D as in definition 6. If all voter's perceived state are in [1, N] - (*$A \cup B \cup C \cup D$*) then* R_b *and* R_s *will be consistent.*

Problems arise when rater's perceived states lie in the critical regions around the $\frac{1}{4}$ and $\frac{3}{4}$ mark of the interval [1,N]. In the case where there are no people with perceived states in this critical region, the binary system aggregation function and the scaled system aggregation function agree. This is due to the fact that the voters in the scaled systems will rate at the endpoints or the midpoint which are exactly equivalent to -1, 0, or 1.

4.3 Implications

This section provides conditions that rater's perceived states must satisfy in order to ensure the binary and scaled metric are consistent. These conditions exploit the difference in mathematical structures of the two systems. In practice, there is no restriction on people's beliefs. People can have a perceived state anywhere in the interval [1,N]. This section illustrates that the two systems

do not necessarily utilize individual's votes in the same way, only in very special cases do they do so. A subsequent question is then to determine the frequency of when the systems disagree and agree.

5 Frequency of Differences Occurring between the Rating Systems

Because of the unsettling notion that metrics are not consistent, questions arise about how often problems may occur. As described in Section 2 above, for every transaction that takes place, there exists a true state TS $\in$ [1, N]. Each individual forms an opinion about TS and obtains a perceived state PS $\in$ [1,N]. The previous sections characterize the differences among the systems in terms of regions in which perceived states lie. This means that in order to discuss how often inconsistencies among systems occur, one must define the probability an individual will have perceived states in certain regions. Using standard assumptions, the distribution of PS is assumed to be normally distributed and centered at TS. This means the majority of people rate close to the true state and justifying for the distribution to be centered around the true state. The farther one's opinion is from TS in the interval [1, N], the more unlikely it is to occur. For example, suppose k people rate a seller's performance in an on-line market. This seller sends a quality product however does not send it to the buyers by the expected date. The true state of the transaction is therefore poor. Suppose the true state of the transaction is approximately 1.5. The k buyers perceive this state differently, however, under the normality assumption, most are around 1.5. Their perceived state distribution will be centered at 1.5 and bell-shaped distributed. If the buyer had sent the quality product on time, then the true state of the transaction may have been near 4.5 on a 1-5 scale. In this case, the k people's perceived state would still be bell-shaped but now centered around 4.5. In other words, the center of the distribution changes in the interval [1,N] but the distribution of opinions remains bell-shaped. Under this normality assumption, the following sections describe the frequency of the differences occurring.

5.1 Frequency of Differences Occurring at the Individual Rater Level

Let f(x) denote the normal distribution function of people's perceived scores in the interval [1,N] that is centered about TS with variance σ^2. The probability of the systems being equivalent is equal to the area under the distribution curve f(x) for the regions of [1, N] on which the two systems are the same.

Theorem 10 *If N is odd and N-1 is divisible by 4, the probability that the binary system and scaled system equals*

$$[\int_{-\infty}^{\frac{1+\frac{N+1}{2}}{2}-\frac{1}{2}} f(x)dx] + [\int_{-\infty}^{\frac{1+\frac{N+1}{2}}{2}+\frac{1}{2}} f(x)dx - \int_{-\infty}^{\frac{N+\frac{N+1}{2}}{2}-\frac{1}{2}}] + [1 - \int_{-\infty}^{\frac{N+\frac{N+1}{2}}{2}+\frac{1}{2}} f(x)dx]$$

where $f(x)$ is the normal p.d.f. of the distribution of perceived scores with the mean TS and standard deviation equals to σ.

Theorem 11 *If N is even and divisible by 4, the probability that the binary system and scaled system equals*

$$[\int_{-\infty}^{\frac{1+\frac{N+1}{2}}{2}-\frac{1}{4}} f(x)dx] + [\int_{-\infty}^{\frac{1+\frac{N+1}{2}}{2}+\frac{3}{4}} f(x)dx - \int_{-\infty}^{\frac{N+\frac{N+1}{2}}{2}-\frac{3}{4}}] + [1 - \int_{-\infty}^{\frac{N+\frac{N+1}{2}}{2}+\frac{1}{4}} f(x)dx]$$

where $f(x)$ is the normal p.d.f. of the distribution of perceived scores with the mean TS and standard deviation equals to σ. If N is even and not divisible by 4, the probability that the binary system and scaled system equals

$$[\int_{-\infty}^{\frac{1+\frac{N+1}{2}}{2}-\frac{3}{4}} f(x)dx] + [\int_{-\infty}^{\frac{1+\frac{N+1}{2}}{2}+\frac{1}{4}} f(x)dx - \int_{-\infty}^{\frac{N+\frac{N+1}{2}}{2}-\frac{1}{4}}] + [1 - \int_{-\infty}^{\frac{N+\frac{N+1}{2}}{2}+\frac{3}{4}} f(x)dx]$$

where $f(x)$ is the normal p.d.f. of the distribution of perceived scores with the mean TS and standard deviation equals to σ.

5.2 Frequency of Differences Occurring at the Aggregation Level

In order to aggregate k raters individual scores, the binary aggregation function R_b counts the number of people rating at -1, 0, and 1 and averages. Similarly, the scaled aggregation function R_s computes the number of people rating at each int[1,N] and then averages the scores. Depending on where the truth lies, certain rating are more likely than others. In this way, the number of individual reports of a specific value in either system is related to the distribution of perceived values centered around the true state. Let $\%b_i$ denote the percentage of raters rating i where i=-1, 0, 1 and let $\%s_j$ be the percentage of raters rating j where j=1, 2, ..., N.

In the binary system, each value can then be expressed in the following manner:

$$\%b_1=\int_{-\infty}^{-.5} f(x)dx, \quad \%b_2=\int_{-.5}^{.5} f(x)dx, \quad \%b_3=\int_{.5}^{\infty} f(x)dx$$

where f(x)= is the normal distribution function of perceived states centered around TS_b with standard deviation denoted as σ.

In the scaled system,

$$\%s_1=\int_{-\infty}^{1.5} f(y)dx, \quad ..., \quad \%s_k=\int_{\frac{k+(k-1)}{2}}^{\frac{(k+1)+k}{2}} f(y)dx, \quad ...,\%s_N=\int_{\frac{N-(N-1)}{2}}^{\infty} f(y)dx$$

where f(y)=is the normal distribution function of perceived states centered around $TS_s = \frac{1+N}{2} + (\frac{N-1}{2})(TS_b)$ and standard deviation $\sigma_s = \frac{\sigma}{\frac{N-1}{2}}$.

Because a linear transformation between the systems exists, the distributions can also be expressed one in terms of the other. The percentage of -1 ratings in the binary system, can be expressed in terms of the distribution in the scaled system.

$$\%b_1= \int_{-\infty}^{-.5} f(x)dx = s_1 \int_{-\infty}^{1.5} f(y)dy + ... + s_{\frac{1+\frac{(N+1)}{2}}{2}} \int_{(\frac{1+\frac{(N+1)}{2}}{2}-\frac{1}{2})}^{(\frac{1+\frac{(N+1)}{2}}{2}+\frac{1}{2})} f(y)dy$$

The number of neutral 0 rankings is:

$$\%b_2 = \int_{-.5}^{.5} f(x)dx = s_{\frac{1+\frac{(N+1)}{2}}{2}+1} \int_{((\frac{1+\frac{(N+1)}{2}}{2}+1)-\frac{1}{2})}^{((\frac{1+\frac{(N+1)}{2}}{2}+1)+\frac{1}{2})} f(y)dy ... + s_{\frac{N+\frac{1+N}{2}}{2}} \int_{\frac{N+\frac{1+N}{2}}{2}-\frac{1}{2}}^{\frac{N+\frac{1+N}{2}}{2}+\frac{1}{2}} f(y)dy$$

The number of +1 rankings is:

$$\%b_3 = \int_{.5}^{\infty} f(x)dx = s_{\frac{N+\frac{1+N}{2}}{2}+1} \int_{(\frac{N+\frac{1+N}{2}}{2}+1)-\frac{1}{2}}^{(\frac{N+\frac{1+N}{2}}{2}+1)+\frac{1}{2}} f(y)dy + ... + s_N \int_{\frac{(N-1)+N}{2}}^{\infty} f(y)dy$$

Combining all of this information allows for the aggregation functions R_b and R_s to be expressed in terms of the probability distribution of perceived states in each system. The above expressions then quantify how likely it is that the two systems will be consistent.

6 Conclusion

Rating systems measuring quality of products and service are typically used to solve the asymmetric information problem in markets. They summarize public opinion and are widely used to stabilize markets, ensure quality of service, and help people assess situations. A rating system is composed by a metric and an aggregation rule. The results in this paper illustrate how different metrics and aggregation procedures may translate the same opinion to different conclusions. Previous literature has explained these differences as a result of human error. However, this paper shows that even without human error, the inconsistencies among systems still exist. This is due to the mathematical structure of the rating systems themselves.

This work characterizes the differences in the binary and scaled systems in terms of the metrics. First, the manner in which the two metrics represent an individual's opinion was examined. Regions of the interval [1, N] where the two metrics produce consistent representations were defined. In particular, it is shown that the binary system is equivalent to the scaled system only in the case where N is odd and N-1 is divisible by 4. This result provides a way to construct the best possible scaled system. Also, it explains inconsistencies seen in applications solely based on the mathematical structure of the metrics. The implications of the different representations, as well as when and how often the differences occur at the individual level, were presented. Secondly, the effect of

the different aggregation rules in the given metric were examined. Simple tools were found that determine whether the binary and scaled rating systems are reporting the same information. In addition, when and how often the system differ at the aggregation level was also characterized. This paper not only provides a new mathematical explanation of the inconsistencies in rating systems, but also provides new tools to study preference information aggregation in social choice and reputation system design.

7 Proofs

Theorem 1:

Suppose not. Suppose the two metrics are equivalent for perceived scores $\in [\frac{1+\frac{1+N}{2}}{2} - \frac{1}{2}, \frac{1+\frac{1+N}{2}}{2} + \frac{1}{2}]$. If N is odd and N-1 is divisible by 4, then $\frac{1+N}{2}$ is an integer and is even. This means that $\frac{1+\frac{1+N}{2}}{2}$ is an integer as well. Suppose two rates have perceived states in $[\frac{1+\frac{1+N}{2}}{2} - \frac{1}{2}, \frac{1+\frac{1+N}{2}}{2} + \frac{1}{2}]$. Then both raters will rate $\frac{1+\frac{1+N}{2}}{2}$ because this is the closest integer available. Let rater 1 have perceived state $\in [\frac{1+\frac{1+N}{2}}{2} - \frac{1}{2}, \frac{1+\frac{1+N}{2}}{2}]$ and rater 2 have perceived state $\in [\frac{1+\frac{1+N}{2}}{2}, \frac{1+\frac{1+N}{2}}{2} + \frac{1}{2}]$ then in binary system, rater 1 will rate -1 and rater 2 will rate 0. This implies the systems are not equivalent. Same argument holds for $[\frac{\frac{1+N}{2}+N}{2} - \frac{1}{2}, \frac{\frac{1+N}{2}+N}{2} + \frac{1}{2}]$.

Theorem 2:

If N-1 is not divisible by 4, then $\frac{1+\frac{1+N}{2}}{2}$ is not an integer. By definition 2, $\frac{1+\frac{1+N}{2}}{2}$ transformed into the binary metric is -0.5. If two rater rate smaller than $\frac{1+\frac{1+N}{2}}{2}$, then both will rate -1 in the binary. If two raters rate larger than $\frac{1+\frac{1+N}{2}}{2}$, then both raters will rate 0 in the binary metric. Because $\frac{1+\frac{1+N}{2}}{2}$ is not an integer and has the form k.5 where k is an integer, then the closest integers are $\frac{1+\frac{1+N}{2}}{2} - \frac{1}{2}$ and $\frac{1+\frac{1+N}{2}}{2} + \frac{1}{2}$. If two raters rate $\frac{1+\frac{1+N}{2}}{2} - \frac{1}{2}$, then their perceived states will be in $[\frac{1+\frac{1+N}{2}}{2} - 1, \frac{1+\frac{1+N}{2}}{2}]$. Any two raters with perceived states in this interval will rate -1 in binary. If two raters rate $\frac{1+\frac{1+N}{2}}{2} + \frac{1}{2}$, then their perceived states will be in $[\frac{1+\frac{1+N}{2}}{2}, \frac{1+\frac{1+N}{2}}{2} + 1]$. Any two raters with perceived states in this interval will rate 0 in binary. The $\frac{N+\frac{1+N}{2}}{2}$ follows in the exact same manner. Therefore the systems are equivalent everywhere.

Theorem 3:

If N is even and divisible by 4, then $\frac{1+N}{2}$ is not an integer. Specifically, it has the form k.5 where k is an integer. Also, $\frac{1+\frac{1+N}{2}}{2}$ is not an integer and has the form p.75 where p is an integer. This means the closest integers to $\frac{1+\frac{1+N}{2}}{2}$ are $\frac{1+\frac{1+N}{2}}{2} - \frac{3}{4}$ and $\frac{1+\frac{1+N}{2}}{2} + \frac{1}{4}$. Suppose two raters rate $\frac{1+\frac{1+N}{2}}{2} + \frac{1}{4}$, then this means their perceived states are $\in [\frac{1+\frac{1+N}{2}}{2} + \frac{1}{4} - \frac{1}{2}, \frac{1+\frac{1+N}{2}}{2} + \frac{3}{4}]$. Suppose rater 1 has perceived state$\in [\frac{1+\frac{1+N}{2}}{2} + \frac{1}{4} - \frac{1}{2}, \frac{1+\frac{1+N}{2}}{2}]$ and rater 2 has perceived state in $[\frac{1+\frac{1+N}{2}}{2}, \frac{1+\frac{1+N}{2}}{2} + \frac{3}{4}]$, then rater 1 rates -1 in the binary and rater 2 rates 0. This means the systems are not equivalent on the interval. Similarly the systems are not equivalent on the interval $[\frac{\frac{1+N}{2}+N}{2} - \frac{3}{4}, \frac{\frac{1+N}{2}+N}{2} + \frac{1}{4}]$.

If N is even and not divisible by 4, then $\frac{1+\frac{1+N}{2}}{2}$ is not an integer and has the form p.25 where p is an integer. The closest integers is then $\frac{1+\frac{1+N}{2}}{2} - \frac{1}{4}$. Suppose two raters rate $\frac{1+\frac{1+N}{2}}{2} - \frac{1}{4}$, then their perceived states could be anywhere $\in [\frac{1+\frac{1+N}{2}}{2} - \frac{1}{4} - \frac{1}{2}, \frac{1+\frac{1+N}{2}}{2} - \frac{1}{4} + \frac{1}{2}]$. Suppose rater 1 has a perceived state in $[\frac{1+\frac{1+N}{2}}{2} - \frac{3}{4}, \frac{1+\frac{1+N}{2}}{2}]$, then rater 1 rates -1 in the binary metric. Suppose rater 2 has a perceived state in $[\frac{1+\frac{1+N}{2}}{2}, \frac{1+\frac{1+N}{2}}{2} + \frac{1}{4}]$, then rater 2 rates 0 in the binary metric. Therefore the systems are not equivalent on the interval. Similarly the systems are not equivalent on the interval $[\frac{\frac{1+N}{2}+N}{2} - \frac{1}{4}, \frac{\frac{1+N}{2}+N}{2} + \frac{3}{4}]$.

Theorem 4:

Perceived state PS_s can be translated to PS_b in the binary system using definition 2. In addition, PS_s and PS_b are rated as r_s and r_b respectively. By definition 2, r_s can be translated to the binary system, call it r'_{sb}. The difference between the binary rating formed by translating the perceived score and the rating formed by translating the scaled rating provide a measure for the penalty or benefit alloted by the binary system.

Theorem 5:

Combining the results of theorems 1,2, and 3, the only scaled metric that is equivalent to the binary metric is the case where N is odd and N-1 is not divisible by 4. Therefore, this scaled metric represents opinions in the same manner as the binary system.

Theorem 6:

Suppose N-1 is not divisible by 4 and |A|=|B|=k where k is an integer. Then those k raters with

perceived states in A will rate $\frac{1+\frac{1+N}{2}}{2} - \frac{1}{2}$ since it is the closest integer available while the k raters in perceived states in B will rate $\frac{1+\frac{1+N}{2}}{2} + \frac{1}{2}$ since that is the closest integer available to them. Then the aggregation function $R_s = \frac{k(\frac{1+\frac{1+N}{2}}{2} - \frac{1}{2}) + k(\frac{1+\frac{1+N}{2}}{2} + \frac{1}{2})}{2k} = \frac{1+\frac{1+N}{2}}{2}$. Using definition 2, this value is equal to $\frac{-1}{2}$ in the binary metric. Now, the k raters with perceived states in region A will rate -1 in the binary metric while the k raters with perceived states in region B will rate 0. This means that $R_b = \frac{k(-1)+k(0)}{2k} = \frac{-1}{2}$. Therefore the aggregation functions are consistent. Similarly for regions C and D.

Suppose N-1 is divisible by 4 and |A|=|B|=k where k is an integer. Then the k raters in both regions will rate $\frac{1+\frac{1+N}{2}}{2}$ since this is the closest integer available to them. Then the aggregation function $R_s = \frac{k(\frac{1+\frac{1+N}{2}}{2}) + k(\frac{1+\frac{1+N}{2}}{2})}{2k} = \frac{1+\frac{1+N}{2}}{2}$. Using definition 2, this value is equal to $\frac{-1}{2}$ in the binary metric. Now, the k raters with perceived states in region A will rate -1 in the binary metric while the k raters with perceived states in region B will rate 0. This means that $R_b = \frac{k(-1)+k(0)}{2k} = \frac{-1}{2}$. Therefore the aggregation functions are consistent. Similarly for regions C and D.

Theorem 7:

Suppose N-1 is not divisible by 4 and |A|=|D|=k where k is an integer. Then those k raters with perceived states in A will rate $\frac{1+\frac{1+N}{2}}{2} - \frac{1}{2}$ since it is the closest integer available while the k raters in perceived states in D will rate $\frac{N+\frac{1+N}{2}}{2} + \frac{1}{2}$ since that is the closest integer available to them. Then the aggregation function $R_s = \frac{k(\frac{1+\frac{1+N}{2}}{2} - \frac{1}{2}) + k(\frac{N+\frac{1+N}{2}}{2} + \frac{1}{2})}{2k} = \frac{1+N}{2}2$. Using definition 2, this value is equal to 0 in the binary metric. Now, the k raters with perceived states in region A will rate -1 in the binary metric while the k raters with perceived states in region D will rate 1. This means that $R_b = \frac{k(-1)+k(1)}{2k} = 0$. Therefore the aggregation functions are consistent. Similarly for regions B and C.

Suppose N-1 is divisible by 4 and |A|=|D|=k where k is an integer. Then the k raters in A will rate $\frac{1+\frac{1+N}{2}}{2}$ since this is the closest integer available while the k raters in D will rate $\frac{N+\frac{1+N}{2}}{2}$. Then the aggregation function $R_s = \frac{k(\frac{1+\frac{1+N}{2}}{2}) + k(\frac{N+\frac{1+N}{2}}{2})}{2k} = \frac{1+N}{2}$. Using definition 2, this value is equal to 0 in the binary metric. Now, the k raters with perceived states in region A will rate -1 in the binary metric while the k raters with perceived states in region D will rate 1. This means that $R_b =$

$\frac{k(-1)+k(1)}{2k}$= 0. Therefore the aggregation functions are consistent. Similarly for regions B and C.

Theorem 8:

Suppose N is divisible by 4 and |A|=|D|=k where k is an integer. Then the k raters in A will rate $\frac{1+\frac{1+N}{2}}{2}$ since this is the closest integer available while the k raters in D will rate $\frac{N+\frac{1+N}{2}}{2}$. Then the aggregation function R_s=$\frac{k(\frac{1+\frac{1+N}{2}}{2})+k(\frac{N+\frac{1+N}{2}}{2})}{2k}$ = $\frac{1+N}{2}$. Using definition 2, this value is equal to 0 in the binary metric. Now, the k raters with perceived states in region A will rate -1 in the binary metric while the k raters with perceived states in region D will rate 1. This means that R_b= $\frac{k(-1)+k(1)}{2k}$= 0. Therefore the aggregation functions are consistent. Similarly for regions B and C and when N is greater than 2 and not divisible by 4.

Theorem 9:

This follows directly from theorems 6,7, and 8. Since no matter N is even or odd, A,B,C,D are the problematic regions that cause inconsistency.

Theorems 10 and 11:

This follows directly from the definition the normal probability distribution and theorems 1, 2, and 3.

References

[1] Alwin, F. D. and Krosnick, A. J. (1985). The Measurement of Values in Surveys: A Comparison of Ratings and Rankings, *The Public Opinion Quarterly*, **49** (4), 535-552

[2] Arrow, K. J. (1950). A difficulty in the Concept of Social Welfare, *Journal of Political Economy*, **58** (4), 469-84.

[3] Arrow, K. J. (1963). *Social Choice and Individual Values*, John Wiley Sons, Inc., New York, London, Sydney.

[4] Cleveland, J. N. and Murphy, K. R., (1992). Analyzing performance appraisal as goal-directed behavior. In G. Ferris K. Rowland (Eds.), *Research in personnel and human resources man-*

agement (Vol. 10, pp. 121185). Greenwich, CT: JAI Press.

[5] Dellarocas, C. (2003). The digitalization of the Word of Mouth: Promise and Challenges of Online Feedback Systems, *Management Science*, **49**, 1407-1424.

[6] C. Dellarocas, F. Dini, and G. Spagnolo. (2006) Designing reputation (feedback) mechanisms. In *Handbook of Procurement*, Cambridge University Press, 2006.

[7] Dellarocas, C and Wood, C. A. The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. *Management Science*, Forthcoming.

[8] Droba, D. D. (1931). Methods Used for Measuring Public Opinion, *The American Journal of Sociology*, **73** (3), 410-423.

[9] E A. Greenleaf, B. Bickart, and E. A. Yorkston, (1999), How Response Styles Weaken Correlations from Rating Scale Surveys, Working paper, Stern School of Business, New York

[10] R. Jurca and B. Faltings. (2005) Enforcing trusthful strategies in incentive compatible reputation mechansims. *Internet and Network Economics*, Lecture Notes in Computer Science, Volume 3828, pp. 268-277

[11] Landy, F. J. and Farr, J. L., (1980), Performance Rating, *Psychological Bulletin*, **87** (1), 72-107.

[12] Lehmann, D. and Hulbert, J., (1972). Are Three-Point Scales Always Good Enough?, *Journal of Marketing Research* **9**, (4), 444-446.

[13] Li, L. I., Reputation, Trust, and Rebates: How Online Markets Can Improve Their Feedback Mechanisms. (October 27, 2006). Institute for Mathematical Behavioral Sciences. Paper 55. http://repositories.cdlib.org/imbs/55

[14] Li, L. I. and Saari, D. G. (2008). Sens theorem: geometric proof, new interpretations, *Social Choice and Welfare*, Online publication date: 21-Jan-2008.

[15] McCarty, J. and Shrum, L.J., (2000). The Measurement of Personal Values in Survey Research: A Test of Alternative Rating Procedures, *The Public Opinion Quarterly*, **64**, (3), 271-298.

[16] McDaniel, C. and Gates, R., (2004). *Marketing Research*, Wiley; 6 edition

[17] N. Miller, P. Resnick, and R. Zeckhauser. (2005) Eliciting honest feedback: The peer prediction method. *Management Science*, September 2005.

[18] K. R. Murphy, J. N. Cleveland, A. L. Skattebo, and T. B. Kinney, (2004), Raters Who Pursue Different Goals Give Different Ratings, *Journal of Applied Psychology*, **89** (1), 158-164.

[19] F. E. Saal, R. G. Downey, and M. A. Lahey, (1980), Rating the Ratings: Assessing the Psychometric Quality of Rating Data, *Psychological Bulletin*, **88** (2), 413-428.

[20] Saari, D.G. (1999). Explaining all three alternative voting outcomes, *Journal of Economic Theory*, **87**, 313-355

[21] Saari, D.G. (2001a) *Chaotic elections! A mathematician looks at voting Providence*, American Mathematical Society.

[22] Saari, D. G. (2001b). *Decisions and Elections; Explaining the Unexpected*, Cambridge University Press, New York.

[23] N. Schwarz, B. Knuper, H. Hippler, E. Noelle-Neumann, and L. Clark, (1991), Rating Scales: Numeric Values May Change the Meaning of Scale Labels, *The Public Opinion Quarterly*, **55** (4), 570-582.

[24] Sen, A. K. (1970a). *Collective Choice and Social Welfare*, Holden-Day, San Francisco.

[25] Sen, A. K. (1970b). The Impossibility of a Paretian Liberal, *The journal of Political Economy*, **78**(1), 152-57.

[26] Sen, A. K. (1986) Social Choice Theory; in *Handbook of Mathematical Economics, Vol. III;* Ed. K.J. Arrow and M. Intriligator, Amsterdam: North-Holland.

[27] Stevens, S. S. (1946). On the Theory of Scales of Measurement, *Science*, **103** (2684), 677-680