Assessing Team Performance in Information Systems Projects

William J. Tastle
School of Business
Ithaca College
Ithaca, New York 24850
tastle@ithaca.edu

Emil Boasson
Business Information Systems Department
Master of Science in Administration Program
Central Michigan University
Mount Pleasant, Michigan 48858
boass1e@cmich.edu

Mark J. Wierman
Dept of Computer Science
Creighton University
Omaha, Nebraska 68178-2090
mwierman@creighton.edu

Abstract

Ordinal scales are a typical way by which subjective data can be collected and analyzed, and traditional statistics are typically used to analyze the resulting data. This can lead to erroneous conclusions. A set of measures has been developed for ordinal scale calculations that are easy to understand and can be quickly applied in the classroom. This paper uses ordinal measures of consensus, dissent and agreement to illustrate a mechanism by which student team progress can be assessed. The data present in a Likert scale frequency distribution is used to calculate the evidence associated with the individual categories and also the degree of dispersion among the categories. Thus, it can be shown that each ordinal category contains a certain amount of evidence, and the overall frequency distribution contains a specific degree of dispersion. Instructors can use this easily attained information to plot the success of teams and to identify those teams that require additional assistance.

Keywords: ordinal scale, consensus, agreement, dispersion

1. Introduction

One perplexing aspect of assigning students to teams, even if the teams are self-assigned, is the effort that occurs within the team environment when a decision must be made on some critical project component. If one assigns projects to student teams it is apparent (and seemingly universal) that student team agreement is sometimes a formidable task. An added component of

complexity occurs when diversity among the students requires action by the instructor (Caspersz, et al, 2004). The authors present a method by which teams can assess themselves during times of conflict.

Arguably, one of the easiest methods by which the perceptions and attitudes of individuals who form a team can be captured is by means of a Likert scale set of categories. By merely checking a box one can usually quantify one's subjective feelings. These categories are ranked in some form of hierarchical order and are, by definition, called ordinal scales. These ordinal scales of measurement typically consist of "strongly agree," "agree," "neither agree nor disagree," "disagree," and "strongly disagree;" or in judging the temperature of a swimming pool as "very cold," "cold," "cool," "tepid," "warm," "hot," and "very hot." Team assessment of agreement is not limited to the classroom environment, for another set of categories is the terror threat level that now pervades American society: green, blue, yellow, orange, and red. A team composed of experts from various governmental agencies could be assembled to formulate some degree of agreement of threat level. This is analogous of a team in the classroom attempting to agree on some aspect of a group problem, and the evaluatory skills applied to the classroom are expectedly transportable to all other academic and job-related tasks.

One instrument that can be used to collect some kinds of data is called the Likert scale, though variations of this scale exist such as Likert-like, Likert-type, and ordered response scales. Instructors utilize this type of instrument to collect data that cannot be ascertained using traditional measures (Tastle and Wierman 2008), for the data being collected are feelings, perceptions, sensations, emotions, impressions, sentiments, opinions, passions, or the like. Unfortunately, the application of standard statistics to these data can be improper (Cohen, et al, 2000; Jamieson, 2004; Pell, 2005). This paper presents some new ideas that can be used in assessing team agreement through the analysis of ordinal scale data. The new ideas are represented by measures of consensus, dissent, agreement, and disagreement.

2. Literature Review

There is much literature on general team decision making (Algert, 2000; Bellamy, et. al., 1994; Johnson, et. al., 1986; Johnson and Johnson, 2000; Katzenbach and Smith, 1992; Lovata, 1987; Prodanovic and Simonovic, 2003; Scholtes, et. al., 1996). Johnson and Johnson (2000) describe seven methods that a team may use to arrive at a decision: decision made by authority without group discussion, decision made by an expert, decision made by averaging the individuals' opinions, decision made by authority after a group decision, decision by minority fiat, decision by majority vote, and decision made by consensus. In none of these cases is it shown that the proximity to consensus or agreement can be ascertained by the team, let alone an external observer (an instructor in our case), though it is possible that the last method, decision by consensus, comes closest. This method requires that the observer commit possibly long periods of time in an observer role, something that few instructors have available. Further, the presence of an authority figure may induce the Hawthorne effect (Adair, 1984).

Prodanovic and Simonovic (2003) came closest to permitting a calculation, but it did not provide a means by which a trajectory of continuing monitoring could be conducted. We do provide (below) such a mechanism for monitoring progress.

The authors were unable to locate any reference that provides a means by which the proximity to consensus could be measured, though, not being psychologists, it is possible that such a reference exists in spite of our extensive investigation of the literature.

3. The Use of Mathematics in Measures of Ordinal Scales

Stevens (1946) introduced four levels of measurement: nominal, ordinal, interval, and ratio.

Collections of categories used to accumulate data that are without any sense of order are called *nominal* scales. An example of a nominal scale is a listing of the days of the week. There is no order such that one may make the claim that Monday is greater than Friday. They simply represent a label.

Ordinal scales are merely *ordered* categories, but the ordering makes comparisons implicitly possible. Testing the temperature of a cup of tea would permit someone to use the words "cold," "warm," and "hot" as their scale of comparative measure. There is no sense of interval scale in this measure and hence, equations such as "cold" + "warm" = "hot," or the average of "warm" and "hot" is "warm and a half," are impractical and illogical. Likert scales fall into this category of measures. The numbers of choices that may be selected as categories of a Likert scale are virtually without limit, although five or seven category scales are the most prominent. The distance between each category, sometimes referred to as the interval, is incorrectly assumed to be equal, though the names of the categories chosen do attempt to create some sense of quasi-equal interval-ness. Nevertheless, there is no empirical evidence to suggest that the interval between "neutral" and "agree" is equal to the interval between "strongly disagree" and "disagree."

Interval scales possess a definite and fixed interval between consecutive values. Sometimes the Likert scale categories are assigned numbers to create a sense of interval and called Likert-like scales. Assigning numbers to categories does not necessarily create a sense of interval-ness in the mind of the survey taker, but it does make justification for using statistics much easier. *Ratio* scales have an absolute zero base, possess an interval, and implicitly possess order. The number line is a ratio scale, and all mathematical operations can be conducted on such a scale.

The dispersion of values about a central value, i.e., the mean, permits an assessment of the strength of the collective respondents' perceptions without placing a focus on an arbitrary numerical interval assignment. Thus, a collective set of ordinal scale values that yield a narrow dispersion can logically be viewed as possessing a greater agreement than one with a wide dispersion. The logic is identical to that of the standard deviation. As will be shown below, the inverse of the consensus measure, called a measure of dissent, informs the instructor of the sense of dispersion using the commonly understood concept of percentage.

The mean requires a fixed interval and a continuous scale, neither of which are available in an ordinal scale. Therefore, we use a measure of dispersion that does not require a fixed interval or continuous scale, but does offer a minimum and maximum value. By using a relative measure, such as a percentage, there need not be any agreed-to interval scale; all that needs to be decided is the location of the extreme values associated with the endpoints of the range of categories, in the case of the Likert scale the extrema are "strongly agree" and "strongly disagree." In the case of the terrorist threat levels, the extremes are "red" and "green."

The consensus measure (and its inverse, dissent), along with the agreement measure (and its inverse, disagreement), satisfies the above considerations. It imposes an interval scale upon the ordinal data but is only weakly dependant upon these intervals, and as such, it is a suitable measure of agreement using ordinal scales. The original motivation for the measure was to resolve a problem dealing with group decision-making dynamics, and was thus named the consensus measure.

4. Discussion

It is necessary to agree on some basic concepts before addressing the individual measures. Hence, within the view of an ordinal scale, say a five-category Likert scale, we define consensus as all survey participants (say, n participants) individually and independently choosing the same ordinal category be it "agree," disagree," or any other category value. A complete *lack* of consensus occurs when ½ n participants take the position of "strongly agree" and the remaining ½ n participants take the "strongly disagree" category. Without any other information present we define the consensus for this latter situation to be zero. Like a legislative body evenly split along Party lines, no consensus can be derived from this extreme situation. Thus, a team discussion in which the participants take opposite extreme views are likely to result in no chance of consensus. The proofs for these measures, as well as various applications of them, are available from a number of references (withheld – to be included in the final version).

The complement to consensus is dissent, and it is the above situation that causes dissent to be maximized, that is, to evaluate to one, for there is maximal strife within the body. If only one person leaves either extreme and moves to any other category, consensus increases (becomes greater than zero), and dissent decreases (is less than one). It is interesting to note that when n is odd, consensus cannot equal zero and dissent cannot equal 1 for one extreme will have one extra datum.

Consensus and dissent depend on some average value to determine their value: the mean, median, or mode. It is recognized that modes and medians are conceptually logical measures of average when dealing with ordinal scales, but some traditions die very hard deaths, and the mean is not to be outdone. When the authors first created the consensus measure they used the mean as the measure of central tendency, and so it has continued though we do advocate the instructor-selected choice of median or mode in addition to mean.

Agreement is a measure that indicates a specific category given all of the individual categorical data. Thus, any frequency distribution over a set of categories can yield an assignment to each of those categories, and the agreement measure will assemble the available evidence and provide a value that indicates the degree of agreement associated with each category. In this case the individual categories are called "targets" and the consensus measure is calculated against each target. The target that has the highest number is the winner. However, it is possible to have secondary targets receiving a value that is very chose to that of the winning target. In those cases, it is up to the instructor to decide it additional data is necessary or if a set of categories should be chosen as the response. Disagreement is 1-agreement, thereby also providing a value that is contained in the 0 to 1 range. It might be more important for a instructor to focus efforts on those attributes for which there is greatest disagreement. In the case of a team struggling to arrive at a consensus, a target can be calculated for each of the possibilities under consideration and a value between 0 and 1 is evaluated to determine the individual makeup of the group. This is superior than merely taking a tally of the group for once the values are gotten it is simple to calculate a measure of dispersion to guide the team in determining the likelihood of eventually arriving at a consensus. It also provides the instructor with a means by which to determine which of the available categorical options are in greatest contention. By eliminating the category possessing the least amount of support the selection set can be reduced, thereby possibly hastening arrival at an ideal consensus.

5. Consensus, Dissent and Agreement

For the definition of measures of consensus, dissent and agreement (Tastle and Wierman, 2005), let *X* be a discrete random variable of size *n* with probability distribution $p(X)$ so that $p_i = P(X_i)$ for *i=1* to *n*. As usual μ_X is the mean of X and $d_X = X_{\max} - X_{\min}$ is the width of *X*. Finally let $d_i = |X_i - \mu_X|$ be the absolute deviation of *X* from the mean. The Consensus, *Cns(X)*, is then defined to be:

$$Cns(\mathbf{X}) = 1 + \sum_{i=1}^{n} p_i \log_2 \left(1 - \frac{|X_i - \mu|}{d_X} \right) \qquad (1)$$

The mirror image of consensus is dissention and has the following form:

$$Dnt(\mathbf{X}) = -\sum_{i=1}^{n} p_i \log_2 \left(1 - \frac{|X_i - \mu|}{d_X} \right) \qquad (2)$$

In other words, *Cns = 1 – Dnt* and *Dnt = 1 – Cns*. One of the interpretations of the dissent measure is that of dispersion. If the frequency distribution is balanced on the extreme categories of the Likert scale, for example at *strongly agree* and *strongly disagree*, the dispersion is maximized at 1 (and the consensus is zero). As the frequency distribution approaches the assignment of all probability to a single category, the dispersion approaches 0 (and the consensus approaches one). This is the essence of the consensus measure: the more the assignments are tightly clustered around one category, the higher the consensus and the less the dissent. This dispersion is always a value in the unit interval, [0..1].

Consensus (Equation 1) can become agreement (Equation 3) when the mean μ_X is replaced with some target value, τ, and

we divide by twice the width, $2d_x$, in the denominator. The target, τ, is usually some desired value identified by the experimenter. For our purposes let us assume that we want to identify the degree to which each team member is supportive of some particular category. We thus want the desired response to be *Strongly Agree*. Since that is the first category in our Likert item, it is assigned a numerical value of 1. Hence, in response to some declarative statement like "The most critical problem is that of a lack of product awareness," we desire for our team of students to strongly agree with this statement, i.e., the target is τ=1. This measure is called *agreement* to distinguish it from measures that use an unspecified target such as the mean, median, or mode. Equation 3 shows τ in place of μ and an expanded width. Doubling the width prevents the equation from exploding when extreme values are reflected in the frequency distribution. We have found the agreement function to work especially well in practice and, for this current work, have limited ourselves to the $2d_x$ denominator. For the most part, either consensus or agreement will work very well, but it is necessary to be consistent in their use. It should also be mentioned that consensus, dissent, and agreement are invariant with respect to linear transformations of the random variable *X*.

$$Agr(X \mid \tau) = 1 + \sum_{i=1}^{n} p_i \log_2 \left(1 - \frac{|X_i - \tau|}{2d_X} \right) \qquad (3)$$

While targeting provides a novel way of measuring distance from a desired goal, it assumes that all elements of the assessment are equally important. That, however, may be a false supposition.

6. Example

Suppose we have a team of five students involved in some group problem, perhaps modeling a complex business system, and let us further suppose that a statement has been raised with respect to the problem, perhaps one as simple as "XYZ is the most critical aspect of our problem." The student team members can select one of five Likert categories (strongly agree [SA], agree [A], neither agree nor disagree, or neutral [N], disagree [D], or strongly disagree [SD]). The following examples illustrate some possible frequency distributions.

Figure 1 (see Appendix A) shows a simple distribution in which every team member, except one, has selected an different category. Only SD remains unselected and A is twice selected. Using equation 3 and substituting 1, 2, 3, 4, and 5 as target values for SA, A, N, D, and SD, agreement (agt) values are derived at the bottom of figure 1. Note that the Agree category has evidence of support at the 84.0% level of agreement. This is no surprise given that the A category contains the largest number. The degree of dispersion over all of the categories is only 0.366, or 36.6%. This is interpreted in the same manner as a standard deviation, except that the measure of dissent provides a value of 0 if all team participants select the same category, and a value of 1 if the team members are exactly maximally divided in their category selection, that is to say, ½ n team members have selected SA and ½ n members have selected SD. In such a case the dissent is maximized at 100%.

In second place however, is category N, not category SA. Category N has 80.1% agreement while category SA has only 70.4% agreement. The reason for the higher agreement in N is that the evidence for neutral is impacted by the adjacent categories. Hence, D and A are "close" to N, thereby adding to the evidence of N. In the case of SA, only A is adjacent, so there is less cumulative evidence to support the placement of SA in second position. The measure of dissent is used to calculate this ordinal dispersion. The log expression in equation 3 captures the order of the categories and gives greater value to contiguous frequencies.

Figure 2 (see Appendix A) illustrates the team members not selecting category N, but a degree of evidence of 76.3% is calculated as the degree of evidence in support of neutral. Simply not selecting a category is not sufficient reason for eliminating it from further consideration, for there exists ample evidence contiguous to the category that it may play a more dominant role that category SA, D, or SD (in this example). Note that the dispersion is 53.2%. We read figure

2 as having a target value of A at 79.5% agreement, but with a dispersion of 53.2%.

Figure 3 (see Appendix A) shows each team member selecting a different category. The evidence associated with the contiguous categories increases the influence each category possesses. Hence, category N has support from A and D thereby giving it a 75.7% degree of evidence in its favor but with an even dispersion of 56.6%. This level of dispersion is greater than figure 2, but less than figure 1.

Lastly, figure 4 (see Appendix A) shows all team members selecting category SA. With the evidence that each category possess, the remaining categories also possess some degree of evidence except for SD, the category that is furthest away. The target of SA is 100% with a dispersion of 0%. This is a maximally strong, and ideal, value for it shows that all team members are in complete agreement.

7. Conclusions

A new method is introduced that can assist the instructor in measuring the on-going performance of teams. This can give him/her an opportunity to direct limited time resources to teams that are more in need of assistance. The calculations can be easily programmed into a spreadsheet and used at will. Students can make as many assessments of performance as they wish, and instructors can require an assessment every x minutes, tabulate the results, and turn in for review. One of the authors has used this method with considerable success. While standard deviations are a traditional measure of variance, the new measure of dispersion (dissent) places the degree of variance between 0 and 1 and hence, can be interpreted as a percentage of dispersion. For the typical student, this is far more understandable than the traditional statistic.

8. References Cited

Algert, N.E. (2000). The Center for Change and Conflict Resolution. (979)775–5335, cccr@bigfoot.com.

Caspersz, D., J. Skene, M. Wu, and M. Boland, 2004, "An approach to managing diversity in student team projects." In *Seeking Educational Excellence.* Proceedings of the 13th Annual Teaching Learning Forum, 9-10 February 2004. Perth: Murdoch University.

G. Adair (1984) "The Hawthorne effect: A reconsideration of the methodological artifact" *Journal of Appl. Psychology* **69** (2), 334-345

Johnson, D.W., and Johnson, F.P. (2000). *Joining together: Group theory and group skills,* 7th ed. Boston: Allyn and Bacon.

Johnson, D.W., Johnson, R.T., and Holubec, E.J., 1986. *Circles of Learning: Cooperation in the Classroom*, rev. ed. Edina, MN: Interaction Book Co.

Katzenbach, J.R., and Smith, D.K., 1992. *Wisdom of Teams*. Boston (Harvard Business School Press).

Lovata, L. M. (1987). Behavioral theories relating to the design of information systems. MIS Quarterly, 11(2), 147-149.

Prodanovic, P. and S. P. Simonovic (2003), "Fuzzy comprovise programming for Group decision making." IEEE Transactions of Systems, Man and Cybernetics, Part A, 33(3), pp. 358-365.

Scholtes, P.R., Joiner, B.L., Streibel, B.J., and Mann, D. (1996). *The Team Handbook*, 2d ed., Oriel, Inc.

Tastle, W.J. and M. J. Wierman, 2007, "Consensus and dissention: a measure of ordinal dispersion." Int'l J of Approximate Reasoning, 45, pp. 531-545.

Appendix A

	Likert Categories				
	SA	**A**	**N**	**D**	**SD**
Student	**1**	**2**	**3**	**4**	**5**
1	0	**1**	0	0	0
2	0	0	0	**1**	0
3	0	0	**1**	0	0
4	0	**1**	0	0	0
5	**1**	0	0	0	0
Total	**1**	**2**	**1**	**1**	**0**
Agt	**0.704**	**0.840**	**0.801**	**0.660**	**0.407**

Figure 1. Dispersion, calculated by the dissent measure, is 0.366.

	Likert Categories				
	SA	**A**	**N**	**D**	**SD**
Student	**1**	**2**	**3**	**4**	**5**
1	**1**	0	0	0	0
2	0	**1**	0	0	0
3	0	0	**1**	0	0
4	0	0	0	**1**	0
5	0	0	0	0	**1**
Total	**1**	**1**	**1**	**1**	**1**
Agt	**0.543**	**0.704**	**0.757**	**0.704**	**0.543**

Figure 3. Dissent (dispersion) is 0.566.

	Likert Categories				
	SA	**A**	**N**	**D**	**SD**
Student	**1**	2	3	4	5
1	**1**	0	0	0	0
2	**1**	0	0	0	0
3	**1**	0	0	0	0
4	**1**	0	0	0	0
5	**1**	0	0	0	0
Total	**5**	**0**	**0**	**0**	**0**
Agt	**1.000**	**0.807**	**0.585**	**0.322**	**0.000**

Figure 4. Dissent (dispersion) is 0.

	Likert Categories				
	SA	**A**	**N**	**D**	**SD**
Student	**1**	**2**	**3**	**4**	**5**
1	0	**1**	0	0	0
2	0	0	0	**1**	0
3	0	0	0	**1**	0
4	0	**1**	0	0	0
5	**1**	0	0	0	0
Total	**1**	**2**	**0**	**2**	**0**
Agt	**0.652**	**0.795**	**0.763**	**0.698**	**0.452**

Figure 2. Dissent (dispersion) is 0.532.