Measuring Faculty Instructional Performance

William J. Tastle
tastle@ithaca.edu
Department of Management, Ithaca College
Ithaca, New York 14850, USA

Abstract

The tenure or promotion of faculty are frequently dependent on the analytical outcome of student end-of-semester evaluation instruments. Faculty usually use the data to make adjustments in their classes, but deans and chairs use the data to determine which faculty are performing "below" average. The statistical measure used is typically the *mean*. The *mean* is invalid for ordinal scale items (like Likert scales), but the consensus, dissent, and agreement measures offer a more intuitive view of data and eliminate the need for the incorrect and morale damaging designation of substandard performance. Successful application of these new measures can permit faculty to measure the perceptions of their students with of colleagues in other schools such that appropriate mentoring can occur.

Keywords: Faculty assessment, visual analog scale, Likert scale, consensus, agreement measure, student course evaluation

1. INTRODUCTION

University teaching supposedly benefits from the solicitation of end-of-semester student evaluations of instructor courses which are analyzed, reflected upon, and results in some consequent action. These reviews have been called course assessments, course evaluations, instructor evaluations, instructor assessment, student reaction forms, etc. Some Deans (and Chairs) use the evaluations to delineate a line separating those faculty who are deemed to have performed above the 'average' from those who have performed below 'average.' Tenure, promotion and annual financial increments typically depend on the Dean's analysis.

We show that the usual analysis of faculty assessment can be inaccurate at best, and invalid at worse.

The instruments themselves vary greatly in question, style, and scale. We focus on the survey scale. They can be classified into following: the visual analogue scale (Cox, et al, 1990), graphic rating scale, slider scales and radio button scales (Funke and Raipps, 2008), discrete visual analogue scales (Uebersax, 2006), and the ever-present Likert scales (Cox, et al, 1990; Uebersax, 2006). Because of space limitations this paper deals exclusively with the Likert scale.

Motivation

In an unscientific poll of regular ISECON attendees over the period 2000-2008 it has become evident that no two evaluation instruments are identical, and the method of analysis is typically little more than the application of some standard, low-level statistical measure. The usual statistic used in the analysis of these surveys is the *mean*, and it is common for no distinction to be made for the different rigors inherent among the various disciplines in schools of business. The disciplines existing in schools of computer science or information systems are relatively homogeneous, but in other schools it is not uncommon for the technical domains to be lumped together with non-technical domains. Hence, IS/IT is measured against management, finance, marketing, accounting, etc. While no claim is being

made to classify disciplines by some 'degree of difficulty' criteria, there are significant differences.

Furthermore, differences in evaluation instruments makes comparisons of overall teaching performance among different institutions difficult if not impossible. Such comparisons would be useful in identifying institutions and programs which stand out in overall teaching effectiveness so that other institutions might forge linkages with faculty from those institutions for the purpose of engaging in instructional and course improvement mentoring. While all professional educators seek to improve their techniques, it is typically done within one's institution. The method discussed in this paper could permit faculty mentoring across institutions.

Within a single school or program, faculty comparisons by administrators (be they deans or chairs) for purposes of promotion or merit increase should be made between similar domains, or so it would seem. Based upon a preliminary investigation with limited access to faculty evaluation summaries suggests that "average" performance differs by instructional domain, though this research could not be properly evaluated due to issues of instructional confidentiality. For purposes of this study we shall assume that there is a need to distinguish between faculty instructional capabilities (*inter-faculty evaluations*) in the home institution, and also a desire to evaluate instructional performance among a set of pre-selected institutions (*intra-institutional evaluations*). Further, given that no two institutions seem to use the same evaluation instrument there needs to be a method by which logical comparisons can be made especially since **many** (perhaps **all** is a better word) institutions have at least some Likert scale items on their evaluation instrument.

This paper examines the use of the *mean* in evaluating a typical evaluation instrument, discusses the limitations and invalid application of the *mean* to certain scales, and discusses a much more sophisticated method of analysis that provides a different view of the data especially when non-interval scale survey questions/items are solicited. No effort is made to discuss discipline ranking.

2. THE TYPICAL *MEAN*

Evaluation instruments vary to the extreme: there are those composed of two or three open-ended items into which the student is expected to comment on his/her perception of instructor teaching competence, and extends to those instruments composed of dozens of items that involve questions of course content, course delivery, instructor preparation, and instructor competence. Undergraduate students are (in some cases) permitted, and encouraged, to render a verdict on the suitability of content in certain courses. The logic of even asking undergraduates that kind of question is beyond the scope of this paper but does lend credence to the need to develop a set of universally accepted assessment criteria. Perhaps this paper will motivate such a future undertaking.

For the sake of this paper we shall assume a relatively simple set of *n* Likert scale statements (called Likert *items*). It is necessary to keep *n* small, but not too small for purposes of realism. Hence, we have arbitrary assigned $n = 7$. Each of the 7 statements is followed by a group of categories from which the student is to make a selection. For example, one of the seven Likert items might be "This course is academically challenging" and the categories are *strongly agree*, *agree*, *neutral*, *disagree*, and *strongly disagree*, referred to throughout this paper as SA, A, N, D, and SD, respectively.

The analysis can be longitudinal in the sense that data collected over time can be compared for the purpose of

- identifying individual instructor trends over time, or
- between specific individuals or groups of individuals to determine an average-ness of instruction (inter-faculty or intra-institutional evaluation).

Seeking a mean value necessitates the understanding that exactly half of the faculty will be ***below average*** and that these faculty so stigmatized may suffer professional and financial damage. It can be argued that such a designation can be detrimental to an individual's career, but we digress.

The *mean* (also called the *arithmetic mean*) is the sum of a list of numbers divided by the number of items in the list. It can be further distinguished as a population or sample mean. While it is the most commonly-used type of average, *median* and *mode* are also types of average. Furthermore, other kinds of "means" have been defined: generalized mean, generalized f-mean, harmonic mean, arithmetic-geometric mean, and various weighted means. The characteristic that is common to all these "mean" variations is the need to quantify some sense of central tendency. We shall confine our discussion to the arithmetic mean for it seems to be the most common form of central tendency used in the analysis of faculty statements based on the unscientific solicitation of ISECON attendees.

The arithmetic mean (henceforth called the *mean* unless there is a possibility of confusion) is greatly influenced by outliers. Skewed distributions, in particular, may cause the calculation of a *mean* that does not capture one's notion of "middle." When one is using a Likert scale to calculate a *mean*, the problem is further accentuated. A short discussion of scale is necessary.

3. SCALES

The theory of scale types (Stevens, 1946) is referenced in virtually all statistics texts. Simply stated, all measurement can be conducted using four different types of scales: "nominal,", "ordinal," "interval," and "ratio."

The *nominal scale* uses codes assigned to objects as labels. For example, rocks can be generally classified into three categories: (1) sedimentary, (2) igneous, and (3) metamorphic. The numbers leading the above rock type is used merely for labeling purposes and has no inherent quality with respect to any organization of the rock types; hence the list of categories could just have easily been (1) igneous, (2) metamorphic, and (3) sedimentary. For this scale valid operations include equivalence (a binary relation on a set that specifies how to partition the set into subsets) and set membership. It is easy to see that there is no meaning that can be associated with the "average of (1) sedimentary and (2) igneous" being (1+2)/2 = 1.5. The central tendency of a nominal scale is represented only by its *mode*.

The *ordinal scale* represents the ranking of objects into some inherent order. The ranking of two items is either "ranked higher than," "ranked lower than," or ranked equal to." There does not exist any natural interval between these categories though one might be tempted to arbitrarily assign numeric values to the categories to "force" a sense of interval. Those forced intervals are invalid in the absence of some criteria that proves, or at least illustrates, a convincing argument.

It is not logical to order the above rock types, but the hardness of minerals (which comprise the rocks) can be ordered. Such a scale is called the Mohs scale of hardness and is explained in any elementary textbook on geology or earth science. In short, one mineral is harder than another if one can scratch the other. Thus, minerals naturally fall into an ordinal listing based on their ability to scratch all the minerals ranked lower. The central tendency of an ordinal scale can be represented by either its *mode* or *median*.

The *interval scale* is distinguished by the presence of a uniform interval. A common example of an interval scale is the Celsius scale in which the unit of measurement is 1/100 of the difference between the melting temperature and boiling temperature of water at sea level. The Fahrenheit scale also contains an interval, though different from the Celsius scale. Regardless how the interval is defined, it is uniform. If there exists a "zero point" on the interval scale, it is arbitrary. In the case of the Celsius scale, zero degrees is defined as the freezing point of water, but it might have been defined as the freezing point of, say CO_2. The central tendency on an interval scale is represented by the *mode*, *median* or *mean*.

The *ratio scale* is an interval scale with the addition of a non-arbitrary zero point. In the case of a variable called temperature the zero point would represent the absence of heat. Such a scale exists and is called the Kelvin scale. The non-arbitrary zero point is absolute zero. All statistical measures can be used at the ratio scale level. The central tendency of a variable on the ratio scale can be the *mode*, *median*, *mean*, and also the *geometric* or *harmonic mean*. See the Ap-

pendix for a summarization of scale types in Table 1.

4. MISUSE OF SCALES

We return to the purpose of this study: to determine a means by which different instructor evaluation instruments, or the summaries of instructional assessments, can be compared. It is typical for instruments to be composed of a set of Likert scale attributes in which the student is requested to identify his/her degree of belief in the attribute by selecting one category from a list of categories such as strongly agree with the attribute, agree with it, disagree, strongly disagree, or hold no particular belief in the attribute. Using the above attribute of "This is an academically challenging course" one would like the students to select "strongly agree." These categories are discrete, but inherently ordered. One can order them as {SA, A, N, D, SD} or {SD, D, N, A, SA}. No other order is possible. Such an ordering is called a *strict total* order and is defined as: $a < b$ if and only if $a \leq b$ and $a \neq b$. If the ordering is not *strict total* then it is possible for a = b, or in words, "strongly agree" is perceived the same as "agree." The point of the Likert scale is to remove this equivalence by selecting Likert categories that implicitly convey a sense of unambiguous ordering.

It is typical to analyze a set of Likert categories by assigning a numeric label to each. Thus, "SA" may be assigned a label of "1", "A" a label of "2", "N" is "3", and so forth. This is akin to assigning a label of "1" to the ordinal category "warm" and a "2" to the category "hot." One would not then take the average of "warm" and "hot" to be (1 + 2)/2 = 1.5 for that would be equivalent to saying that the average of "warm" and "hot" is "warm-and-a-half." Yet in the evaluation of Likert scale statements it is typical to evaluate the average of SA (assigned a value of 1) and SD (assigned a value of 5) as being N (1 + 5)/2 = 3. The use of the mean is mathematically limited to interval and ratio scales. With respect to a measure eof central tendency the best an ordinal scale can identify is either mode or median.

5. Measures of Consensus and Agreement

It has been previously shown and proven (Tastle and Wierman, 2005a, 2005b, 2005c, 2005d, 2006a, 2006b, 2006c, 2007; Tastle, Russell and Wierman, 2005; Tastle, Wierman and Dumdum, 2005; Wierman and Tastle, 2006) that ordinal scales can be viewed from a perspective other than that of interval statistics. In short, the underlying concept involving the creation of *consensus theory* follows:

When dealing with a set of ordinal scale categories, it seems odd to rely on the simple calculation of a *mean*. Besides the obvious problem that the mean requires the presence of a uniform interval, the logic of the resulting value is highly suspect. We take an extreme example to show this point: Suppose 50% of the survey respondents for a certain question select the *strongly agree* (SA) category and the remaining 50% of respondents select *strongly disagree* (SD) as their response, is it logical to take the mean, *neutral* (N) in this case, and use it to represent the result? Even if the standard deviation is also provided (in this case it is 2) does this provide the proper mental perception of the survey respondents' intent? Statistically, the mean and standard deviation do correctly provide the properly accepted method for understanding a set of numbers, but those numbers must reside on an interval line. No evidence exists to support the claim that Likert categories are interval in nature, though it is not uncommon to see ratio statistics applied to Likert scales associated with medical data (Velleman and Wilkinson, 1993). And the academic debate rages on.

Before proceeding any further, it is necessary to clarify some definitions with respect to Likert scales. First, we distinguish between a "Likert scale" and a "Likert item" in that the scale is the sum of responses on several Likert items. Second, we distinguish among the categories of SA, A, N, D, SD as "Likert categories." Hence the sum of the survey responses results in the *Likert scale*, the individual statements are *Likert items*, and the category choices (SA, A, etc) are *Likert categories*.

Building off the mathematics of the Shannon entropy and assorted fuzzy measures (Klir and Folger, 1988) (e.g., belief measures, necessity measures, possibility measures, and the like) a measure has been developed that assigns a value between 0 and 1 to a Likert item.

Example

Given a set of 10 respondents to a single Likert item, exactly 50% select SA and the remaining 50% select SD, it is apparent that the two groups are maximally apart and hence, at maximum disagreement. There is no consensus *with respect to the entire group*. In this situation, the consensus is defined to be zero.

If the same set of 10 respondents all select the same category for another single Likert item, we can say they are in complete consensus *on that item* and this measure assigns a value of 1 to it.

As the distribution of assignments made by the 10 respondents become randomly distributed across all the Likert categories (e.g., SA = 2, A = 3, N = 3, D = 0, SD = 2) it is reasonable to say that we are unsure as to the degree of consensus on this particular Likert item. Is there an appreciable difference between the above distribution (which we shall write as a 5-tuple in the order of SA, A, N, D, SD as {2, 3, 3, 0, 3}, and a different 5-tuple {3, 2, 3, 3, 0}? Using the mean to determine a central tendency yields values of 2.13 and 2.40, respectively, and both 5-tuples (sometimes called a *quintuple* or *pentuple*) have a median of 2.5. The author argues that little insight is gained by using the mean.

Consensus Measure

The equation for consensus is:

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

where X represents the list of categories (Strongly Agree (SA), Agree (A), Neutral (N), Disagree (D), and Strongly Disagree (SD)), X_i is an element of X, μ_X is the mean of X and d_X is the width of X, $d_X = X_{\max} - X_{\min}$.

It is useful to note that *1 – consensus* is equal to a measure of *dissent* (see Tastle and Wierman (2006c) for details). Hence, then there exists complete consensus (Cns = 1), there is no dissent (Dst = 0). If the group of respondents is split with 50% of the group at each end of the Likert scale (SA and SD), then consensus is zero (Cns = 0) and it is reasonable to expect that the amount of dissent is maximized a 100% (Dst = 1).

Applying the consensus measure to the above 5-tuples ({2, 3, 3, 0, 3} and {3, 2, 3, 3, 0}) we have 0.476 and 0.565, respectively. Converting these to percentages, 47.6% and 56.5% we are now able to say that there is a stronger consensus for the second 5-tuple. In fact, we can say that the second 5-tuple is (0.565 – 0.476 = 0.089) 8.9% higher in consensus over the first 5-tuple.

How Much Consensus is a Consensus?

Is the 56.5% consensus in the previous paragraph sufficient to establish what one would typically accept as indicating a group consensus? We suspect not. At the one end of our consensus range is 100%, and to that it is reasonable to say that it is far more than just a consensus; it is total agreement within the survey respondents (we can refer to the group of respondents as stakeholders). At the 50% level of consensus it is reasonable to say that there still remains more work to do to get the stakeholders to perceive themselves at having arrived at a consensus. Statistically speaking we like to use 95% confidence to show satisfaction with our research.

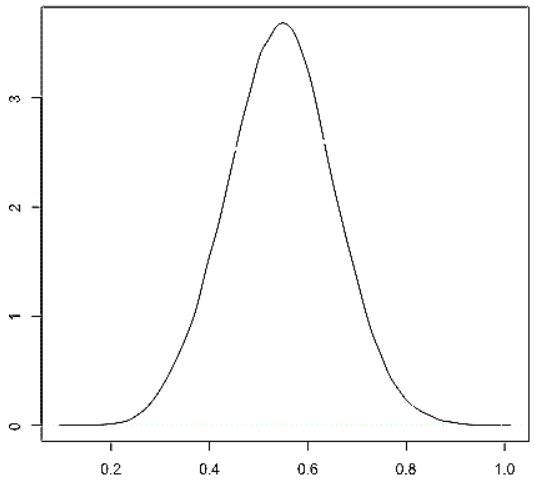

Figure 1 An approximation of consensus. X-axis represents consensus; Y-axis is the density.

Using a bootstrap approximation of 100,000 iterations, and assuming the probabilities are uniform, a distribution results (see figure 1) that has minimum = 0.1242387, 1st quartile = 0.4710029, median = 0.5449004, 3rd quartile = 0.6173651, maximum = 0.9839212, mean = 0.544555, and standard deviation = 0.1088650. It is not unreasonable to say that a 95% confidence interval cutoff will be approximately an 80% consensus. Previously, another author (Salmoni et al, under review) arbitrarily used the 80% consensus as an indication of acceptance. At that time the 80% value was based only on a "hunch," but it is beginning to gather momentum as an appropriate value for accepting the premise that consensus has been reached. More must be done to prove this suspicion (and it is the subject of future research).

Agreement Measure

The desire to *target* the consensus led the authors to examine different expressions for the log term. This has led to the development of an Agreement measure:

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

τ represents the target category such as SA, A, etc. and the denominator is changed to $2d_X$ to reign in the range of the measure.

This measure permits us to measure the overall amount of agreement a group has for a set of Likert items. More to the point, the presence of an agreement 5-tuple moves the eye of the user of the data away from the ambiguous, and frequently invalid, mean (in which half of the faculty will **always** be substandard) to a view that encourages professional educators striving towards excellence in teaching. Comparison with faculty from other disciplines or external to a school does nothing to promote morale or a desire for excellence; the cause around which a dean or chair should rally is *individual faculty improvement*. An example will offer insights into the value of this equation.

Example

Let us assume that we have taken evaluations from a particular class of students at a specific institution spread over three semesters, say an introductory course in IS. The number of students in each class is not constant (few classes have the same number of students each semester) but the course review survey is identical. The following table represents the summarized data:

	SA	A	N	D	SD
Sem1	251	111	10	3	0
Sem2	195	116	31	6	0
Sem3	172	221	57	71	21

Table 2 Frequency distributions for three courses over three semesters. The table is read as 251 semester 1 students selected SA

Let SA be assigned the label of "1", and to A the label of "2", etc. The mean scores for each of the three semesters are 1.373, 1.563, and 2.166. The first two mean scores are quick to be interpreted as being between SA and A, and the third mean is between A and N, but "close" to A. It is easy to suggest that the rating for semester 3 is not as good as the other two semesters.

	Cns	Dst	Mean
Semester 1	80%	20%	1.373
Semester 2	74%	26%	1.563
Semester 3	60%	40%	2.166

Table 3 Consensus (CNS), dissent (Dst), and Mean for each semester.

Using the 80% consensus as previously determined, it is noteworthy that only semester 1 has a frequency distribution that warrants serious attention. The dissent (Dst) percentage (see table 3) is interpreted as the degree of dispersion. Recall, when consensus is 100%, all respondents have chosen the same category (SA, D, etc) and there is 0% dissent. It is evident that little of significance can be said about semesters 2 and 3 and still be within the 95% confidence interval. Using the agreement measure (equation 2) we can easily see that the degree of agreement among the students is beginning to emerge (table 4).

	SA	A	N	D	SD
Semester 1	93%	86%	66%	42%	12%
Semester 2	89%	87%	70%	46%	17%
Semester 3	75%	84%	75%	59%	34%

Table 4 The degree of agreement for each category in each semester.

The interpretation is straightforward: semester 1 has the highest level of agreement in SA and A categories and the least agreement in the N, D, and SD categories. Semester 2 still represents a strong showing in SA and A, but semester 3 shows only a strong agreement in category A. It is interesting to note that agreement is evenly split between SA and N. If figure 4 represents the data from a single faculty member and semester 1 is the most recent semester, then it is apparent that improvements are being made in the classroom. If semester 3 is the most recent, then one must ask what has happened to cause this obvious reduction in student perception. The data acknowledge a potential problem that may be of the faculty member's making, or it might be something out of his/her control (perhaps the class was moved from a computer classroom to one with laptops limited to wireless access). Situations that might be a problem can be easily identified while not embarrassing faculty who are doing very well in their courses.

The presence of a consensus at 80% or greater means that we should have confidence that the agreement 5-tuple accurately represents student intention. If the consensus is less than 80%, there is too much dispersion among the data to make any meaningful conclusion.

6. CONCLUSION

The use of the mean as the primary mechanism for a litmus test of faculty performance is inappropriate at best, and conceptually wrong at worst. Calculations of means are valid only when the data being evaluated are based on an interval or ratio scale. The consensus, dissent and agreement measures, when used in consort provide a much richer, and accurate, view of ordinal data.

The consensus measure gives an accurate indication of overall group support for the categories selected while the measure of dissent is usually interpreted as the degree of dispersion around the consensus. However, consensus gives us no information as to the central tendency, unless the median is used as such. The application of the measure of agreement gives the degree of group support by category that is intuitive, mathematically justifiable, and far more useful.

This method permits different schools to make comparisons of student perceptions with respect to common courses. This is especially relevant if one is attempting to tweak the instruction in core courses of the IS curriculum and one wishes to learn from the successful experiences of other colleagues.

REFERENCES

Cox, C. L., Cowell, J. M., Marion, L. N., and Miller, E. H. (1990), "The Healther Self-Determinism Index for Children," Research in Nursing & Health, 13, pp. 237-246.

Funke, F. and Reipps, U-D. (2008), Difference and Correspondences Between Visual Analogue Scales, Slider Scales and Radio Button Scales in Web Surveys. Available at http://www.frederikfunke.de/papers/. Accessed 6 June 2009.

Klir, G. and Folger T. (1988) "Fuzzy Sets, Uncertainty, and Information," Prentice Hall, Englewood Cliffs, NJ, ISBN 0-13-345984-5.

Salmoni, A. J., Coxall, S., Gonzalez, M., Tastle, W., and Finley, A. (under review) " Defining a postgraduate curriculum in dermatology for general practitioners: a needs analysis using a modified Delphi method."

Stevens, S. S., 1946, "On the theory of scales of measurement. *Science*, 103, pp. 677-680.

Tastle, W. and Wierman, W. (2005a) "Consensus and Dissention: A New Measure of Agreement," North American Fuzzy Information Processing Society (NAFIPS) Conference, Ann Arbor, MI.

Tastle, W. and Wierman, M. (2005b) "Consensus and Dissention: Theory and Properties," North American Fuzzy Information Processing Society (NAFIPS) Conference, Ann Arbor, MI.

Tastle, W. and Wierman, M. (2005c) "A Tool for the Analysis of Ordinal Scale Sate: Measuring Consensus, Agreement, and Dissent," 5th International Conference on Methods and Techniques in Behavioral Research, Wageningen, The Netherlands.

Tastle, W. and Wierman, M. (2005d) "Measuring Consensus in Team Settings Doing Non-routine Knowledge Work," 4th AIS Symposium on Research in Systems Analysis and Design, Univ of Cincinnati [2005], April 23-24, 2005, Univ of Cincinnati.

Tastle, W. and Wierman, W. (2006a) "Adjusting the Consensus Measure to Satisfy Ordinal Scale Arguments," North American Fuzzy Information Processing Society (NAFIPS 2006) Conference, Montreal, Canada.

Tastle, W. and Wierman, M. (2006b) "An Information Theoretic Measure for the Evaluation of Ordinal Scale Data," Journal of Behavior Research Methods, 38(3), pp. 487-494, August 2006.

Tastle, W. and Wierman, M. (2006c) "Consensus and dissention: A measure of ordinal dispersion," International Journal of Approximate Reasoning, 45(3), pp. 531-545.

Tastle, W. and Wierman, M. (2007) "Determining Risk Assessment Using the Weighted Ordinal Agreement Measure," Journal of Homeland Security, http://www. homelandsecurity.org/newjournalArticles / displayArticle2.asp?article=157, June 2007.

Tastle, W., Russell, J. and Wierman, M. (2005) "Ordinal Scales: Using the Consensus Measure to Compare Likert Scale Data," Information Systems Education Conference (ISECON 2005), Columbus, Ohio.

Tastle, W., Wierman M. and Dumdum, U. R. (2005) "Ranking Ordinal Scales Using the Consensus Measure," Int'l Assn of Computer Information Systems Conference (IACIS 2005), Atlanta, GA, pp. 96-102.

Uebersax J. S.(2006) "Likert scales: dispelling the confusion," *Statistical Methods for Rater Agreement* website. 2006. Available at: http://ourworld.compuserve.com/homepages/jsuebersax/likert2.htm. Accessed: *4 June 2009.*

Velleman, P. F. and Wilkinson, L. (1993) "Nominal, Ordinal, Interval, and Ratio Typologies are Misleading. The American Statistician, 47(1), pp. 65-72.

Wierman, M. and Tastle, W. (2006) "Placing the dissonance measure in the context of generalized information theory," In North American Fuzzy Information Processing Society (NAFIPS 2006) Conference, Montreal, Canada.

Appendix

Scale Type	Permissible Statistics	Admissible Scale Trans	Mathematical structure
Nominal (also denoted as categorical or discrete)	Mode, chi square	One to one (equality (=))	Standard set structure (unordered)
Ordinal	Median, percentile	Monotonic increasing (order (<))	Totally ordered set
Interval	Mean, standard deviation, correlation, regression, analysis of variance	Positive liner (affine)	Affine line
Ratio	All statistics permitted for interval scales plus the following: geometric mean, harmonic mean, coefficient of variation, logarithms	Positive similarities (multiplication)	Field

Table 1 Classification of the four different types of scales, Stevens (1946, 1951), taken from Wikipedia. Note that each scale type includes the permissible statistics of the previous types; hence, ordinal statistics include those in the nominal category (mode, chi square).