THE EVALUATION OF PERSONNEL SURVEYS

William Bricken
October 1986

These comments are in reference to the personnel survey that we all recently filled out. My intent is not to analyze the content of the individual questions, but to comment on the overall form of the form. This note discusses what we can learn from the questionnaire and how to learn it. Quite honestly, I don't trust the *third party* that is analyzing the results to do a good job. I hope to be wrong, and will accept appropriate nose-thumbs if I miss the boat.

ASSUMPTION: The information garnered from the questionnaire is in the form of numbers, specifically a three dimensional matrix of person by question by rating, with zeros or ones in each cell.

We have about 100 x 140 x 5 = 70000 bits of information; to make it informative, it needs to be condensed. But first, some comments about the information gathering technique.

OBSERVATION 1: If a personnel questionnaire is anonymous, there is already trouble.

Real opinions are usually voiced by real people. Refinement of a working environment requires dialog, discussion, and negotiation. So let's assume that the questionnaire is setting a ground for a more extensive interaction.

OBSERVATION 2: Human dynamics is poorly represented by a matrix of numbers.

This is an understatement. Opinions are not quantifiable, scalable, or understood through syntactic transformations. So let's assume that the numbers are just setting a ground for some more extensive interaction.

OBSERVATION 3: The entire population of interest is sampled.

Thus, there is no estimating going on. The only problem is to figure out what the pattern of zeros and ones tells us about ADS. Usually, the questions are assumed to generate information. But first, some comments about the other dimensions.

RATING: What does the 1 through 5 scale mean? The *raw data* is binary, we marked one box per question. To boil down the rating dimension to a single number (from 1 to 5) is to make these assumptions:

-- The rating is on a single dimension. Disagree is the exact opposite of agree for the meaning of every question.

-- The rating scale is interval. Strongly agree is exactly twice as strong as agree.

-- The pivot (neutral) is exactly in the center of the scale. Neutrality is on the same continuum as agree and disagree and is half way between.

You might be saying that I'm being a psychometric tightass, which is correct. Scads of social psychological literature shows that these assumptions just aren't true in the domain of human opinions. By permitting *slack* in the form of assumed quantification of belief, we are stripping ourselves of information about humans and defaulting to a mechanical model. To see some of the consequences, consider what happens when scores are summarized across people. The average rating for a particular question might be seen as the company score. But:

-- Do we all interpret each question in the same manner?

-- What does it mean to agree with broadly phrased question?

-- Do we all have the same intent? To provide information, to affect policy, to represent a faction, to embody an ideal?

-- Do we have consistent opinions? What is the effect of recent events on how we feel?

-- Do we all have equal background information, and equal experiences, and needs?

PUNCH LINE 1: Human interpretation of one liners is so diverse that summing across different people introduces only randomness. Scores will drift toward neutral because nothing else makes sense.

What if there's some question that gets a 4, which is not neutral. What do we know then? Still nothing in the way of policy, leadership, decision making. We have described only the communal environment on one question, not our own opinions.

OBSERVATION : Those who have the most deviant opinions will have the most impact on the scores.

In the case of the mean, a 1 might kill of several 4s. The effect is stronger yet when questions are combined.

TECHNICAL OBSERVATION: When combining items (or persons), the weight, or relative effect, of an item is in proportion to its variance, not its mean.

If everyone answers question 42 with a 4, then there is no variation, and no contribution of that question to any factor which is composed of questions. Likewise, if I answer all questions with a 3, then I make no contribution to the variance pool of the replies. That is, people which widely deviant opinions have the most say. (I always vote an extreme when it matters.)

PEOPLE: The second dimension is people. What does william think? Well, we can't tell since we set it up to not have direct information. What about groups of people? Are there factions? Are the new folks different than the old? Well, we have work-title related subgroups, but that is all. However these subgroups reflect only pay scale and not tasks performed. Basically, we're limited to some generic company stereotype.

What kind of information is provided by homogenization? Average information. How many average folks does it take to screw in a light bulb? 3.42

OBSERVATION : The Law of Medium Numbers (thanks to Gerald Weinberg): From 30 to 300 data points, almost anything can happen and it usually does.

OBSERVATION : Homogenization of subpopulations assures that no specific group of opinions is addressed.

Now the hard part:

QUESTIONS: It is usually assumed that questions have some generally agreed upon content. But we were given no map of the territory, no indication of what the questions are getting at. That is to say:

OBSERVATION : People usually have gestalt assessments of a situation. They form a general feeling which dictates replies somewhat regardless of a question's content.

Did you find yourself saying: it's time for a strongly disagree soon.

TECHNICAL OBSERVATION : The only justification for more than a dozen or so questions is replication.

So what questions are intended to go together? What will it mean if I agree with my salary but not my responsibilities. Worse, what will it mean if we all 2.16 agree with salary but 3.09 disagree with responsibility?

There are techniques. A factor analysis over questions is mandatory.

TECHNICAL OBSERVATION : The validity of a set of questionnaire responses is determined by comparison to some criteria.

This is why I what to know the map. Who made up the specific questions, and

for what purpose? Does the sum of (arbitrarily) numbers 9 and 45 and 121 tell us something about morale? Can you justify it? Is there a policy commitment if the threshold of (34 + 68 + 119) is greater than 3.67?

CONFESSION : I took a PhD in Social Psychology and Measurement in the early 70s. There was no symbolic smarts, so we made up questionnaires which asked for ranks from strongly agree to strongly disagree. Then used multivariate analysis to abstract averages. Dropped out because it had nothing to do with social psychology. So I worked for two years constructing and testing selection devices and promotional questionnaires for LA Personnel. Found out that folks with homes tend to stay longer. And that tests, interviews, and questionnaires had nothing to do with performance. Dropped out because that stuff had nothing to do with personnel. So I became a state school teacher and made up tests and questionnaires to find out if the kids learned anything. They learned how to take tests and questionnaires and wondered what it had to do with anything. Of course, test performance is positively related to SAT scores. Dropped out because that had nothing to do with learning. So I began to teach teachers how to deal with groups of people rather than groups of numbers. Lots of the teachers dropped out. I followed. So I went and sat in a forest for many years and learned how to see things (and people) as fractals. And as distinctions. Got excited and enrolled in a PhD in Mathematical Methods of Measurement. Learned that matrices of numbers work really well on machines and have nothing to do with people. Studied kids making mistakes for two years and found that errors are unique. No duplicates, no predictables. My advisor dropped out. So I got this neat symbolic AI job...

SO WHY NUMBERS: It's easier for machine scoring.