Measuring the ‘Unmeasurable’

Jay Alden
July 2, 2007

From time to time, I hear people say something like “You just can’t measure that” or “You really can’t put a numerical score on that.” The point is often made in connection with highly abstract factors and beliefs held by people. What I think these people are really saying is that the factor is too complex to be represented by a single number; too many facets and subordinate qualities are associated with the factor to make a single numerical score meaningful. For example, how would you put a number on the depth of the relationship between a married couple? Yet, I personally believe that you can measure anything. If you’re willing to put the time, effort, and investment in the measurement process, you can come up with a numerical score that can truly represent any factor, no matter how abstract. Wouldn’t *human intelligence* be considered a highly complex and abstract factor with many facets and subordinate qualities? Yet, IQ scores are assigned to most people and these numerical scores seem generally accepted by almost everyone. The Software Engineering Institute of Carnegie-Mellon University assigns software organizations a score from 1 to 5 to represent their *maturity* in software engineering (a score with enormous implications for the organization’s business development capability) and yet this “measure” goes virtually unchallenged. Yes, I’m pretty certain that you can measure anything, so let me describe a few approaches for assigning a numerical score to represent abstract factors using measurement techniques that are considered accurate and meaningful.

1. Find an *Indicator*

The simplest way of attempting to measure an abstraction is to find an easy-to-measure factor that somehow seems to relate directly to the abstract quality. For example, use *number of customer complaints* to represent “customer satisfaction” or use *percentage of reorders* to represent “product quality.” This may be the easiest way to quantify an abstract factor, but the search for such an indicator is usually unsuccessful because there are often conditions in which a single indicator does not validly represent the abstract factor (e.g., it is very difficult for customers to lodge complaints, so even dissatisfied customers don't complain, or you offer the only contractual way of obtaining a product, so customers who feel the product isn't that good, still order it again). Yet, don’t give up on this approach too quickly because if you can find such an indicator, the measurement process is generally quite inexpensive, especially if the organization already measures that indicator for other purposes. If you do unearth a potential indicator, convene a group of people who have insight into the way the measure will be used in the organization, and determine if there is agreement among them that the indicator does, as a general rule, represent the abstract factor of concern. If you do gain consensus, you will have established *face validity* (i.e., scores on the indicator do correspond to the abstract factor, at least on the surface). Say, an organization wishes to determine the degree of “employee satisfaction” in order to decide on whether new human resources initiatives are necessary. The organization might consider using *turnover rate* as an indicator of “employee satisfaction.” If the general consensus is that employees who leave the organization tend to be dissatisfied with its policies (as determined by exit interviews) and employees who remain seem reasonably satisfied (as suggested by observation of their behavior while at work), then the measure of *turnover rate* would be a meaningful indicator of “employee satisfaction.” On the other hand, if people often leave the organization for reasons other than dissatisfaction with its policies while more than a few seemingly unhappy employees remain on the job because of a poor job market, *turnover rate* would probably be an invalid indicator of “employee satisfaction” under these circumstances. A more sophisticated approach to measuring the abstract factor is necessary.

2. Build an *Index*

An “index” is a single score representing some complex factor that is constructed by aggregating the values of several different measures. Sometimes the measures comprising the index are weighted based on their significance to the abstract factor being measured. You’ve probably heard of the *Consumer Price Index* which is a measure of the average price of goods and service purchased by the typical consumer – it includes such components as costs of food and beverages, housing, apparel, transportation, medical care, recreation, education, communication, and energy. Here’s another example of an index:

The *Safety Index* score is used by a global construction company to track the effectiveness of its worldwide safety programs. The index score is composed of four independent measures, each weighted in the calculation of the index to represent its criticality with regard to the concept of “safety.” Suppose that in a given month, world wide on-job injuries in the company resulted in one fatality, seven terminations, 232 lost work days, and 540 restricted work days. These figures would produce a *safety index* score of 1,094.

[(20 X 1) + (10 X 7) + (2 X 232) + (1 X 540) = 1,094]

The index score is tracked month-to-month to assess the impact of its safety program initiatives.

MEASURE		WEIGHT		WEIGHTED SCORE
No. fatalities due to on-job injury	x	20	=	20 x No. FATALITIES
No. terminations due to on-job injury	x	10	=	10 x No. TERMINATIONS
No. lost work days due to on-job injury	x	2	=	2 X No. LOST WORK DAYS
No. restricted work days due to on-job injury	x	1	=	1 X No. RESTRICTED WORK DAYS

SAFETY INDEX = WEIGHTED SUM

One of the benefits of using an index score to represent an abstract factor is that you can drill down to the individual scores comprising the overall measure to identify which ones might be contributing to substandard performance. For example, an unusually high consumer price index might have been caused by a spike in *energy* costs.

The idea when constructing an index is to identify a variety of independent indicators that collectively can represent the abstract factor to be measured. For example, suppose you wished to build a "customer satisfaction index." What indicators tend to suggest satisfaction or dissatisfaction by customers? Several direct indicators might come immediately to mind, such as *number of customer complaints and testimonials, percent of customer reorders or renewals, increase/decrease of reorder size compared to original order, number of cancellations or returns by customers,* and *number of items on backorder*. These indicators of customer satisfaction may be supplemented by some leading indicators typically predictive of customer satisfaction (e.g., *waiting time, product error rate, problem resolution time*). Here's a general process for building an index:

- Find potential indicators: Research the theory and components of the abstract factor being measured using books, journals, and Internet resources and make use of brainstorming sessions involving people knowledgeable in both the factor and business operations. Create a list of potential indicators.

- Select the best indicators: Consider the qualities that would make an effective indicator (e.g., *relevance* to the abstract factor being measured, *importance* to the people who will make use of the index, and *ease* of obtaining data on the indicator). Compare the potential indicators to these qualities and choose the best ones for inclusion in the index

- Assure a common scale among the selected indicators: Since the scores on the various indicators comprising the index must be added together, they each need to use a common scale for the index to make sense. If the indicators use different scales (e.g., say, some are a numerical count of things while others are scored in dollars), estimate the maximum practical numerical value for each indicator and set that value to a score of "100." In the *customer satisfaction index* example, if the maximum size of a reorder based on experience is around $10,000 and the number of customer complaints in any month is likely not to exceed 12, set both those vales to scores of 100. So, an order of $8,000 would contribute 80 points to the index score while six complaints would contribute 50 points to the index. On the other hand, if all the indicators already use a common scale, then just use their normal scores. \

- Adjust the indictors for "favorability": In some cases, the score on an indicator should increase the measure of the index while in other cases the indicator score should reduce the index measure. For example, a *customer satisfaction index* should increase when the size of a reorder is larger because an increased reorder size is considered more favorable, but should decrease when the number of complaints increases because the larger number is considered unfavorable. So, if the various indicators chosen for an index vary in favorability with regard to the abstract factor being measured by the index (i.e., some are positive and some are negative), the calculation should assign plus and minus signs to the indicators accordingly.

- Weight the indicators appropriately: If one or more of the selected indicators is considered vastly more important with regard to the abstract factor being measured, weighting should be used in the calculation of the index. Recall that weighting was used in the calculation of the safety index where *number of fatalities due to on-job injury* was considered 20 times more significant than *number of restricted work days due to on-job injury*. The choice of weighting is a judgment call as is the choice of the particular multiplying values when weighting is used.

- Validate the index: The work up until now produces a *draft* index. Before putting the measure into actual use, you should check to assure that the index will validly produce scores that represent the abstract factor being measured. Two approaches can be used to assure validity. One is to use *face validit*y as described when attempting to use a single indicator to measure a complex factor. That is, try to gain consensus from a group of people with insight into how the index will be used in the organization that, based on the selection and weighting of indicators, the calculated index score will in fact represent the abstract factor intended to be measured. The other approach is to use *content validity*. In this case, you would need to demonstrate that the selection of indicators comprising the index complies with recognized documented references relating to the factor being measured.

Yes, it is a lot of work to build a valid and functional index that represents a complex factor. But the challenge is to measure this complex factor because it is terribly important to do so. Yet, even when it is necessary to measure the complex factor, you might be unsuccessful in building an index that works. There may not be enough indicators to represent the complex factor validly or it might be entirely impractical to develop indicators that will do the job. You'll need to do something else.

3. Construct an *Instrument*

A measurement *instrument* is a tool constructed to assign a numerical score to a complex factor directly and specifically. The most common types of measurement instruments used in organizational assessment are *questionnaires* and *rating forms*. In some respects, questionnaires and rating forms are very similar to each other, in that both kinds of instruments (1) are composed of multiple items, each assessing a different dimension of the factor being evaluated, and (2) are built and tested to yield a single valid and reliable score derived by averaging or summing the scores on the individual items. On the other hand, rating forms differ from attitude questionnaires in a couple of significant ways:

	QUESTIONNAIRE	RATING FORM
On what basis do the people completing the form make their responses?	Scores assigned by the respondent represent how he or she *feels* about the various characteristics of the entity drawing on their own personal values and beliefs (which may rightfully differ from one respondent to another).	Scores assigned by the respondent represent how well the characteristics of the entity being assessed meet given standards of performance (which are either provided or assumed to be known by the rater).
How is the final score derived?	The final score is derived by averaging the responses of <u>multiple</u> users who legitimately might have different feelings concerning the single entity being assessed based on their own perspectives (e.g., a sample of customers express their personal satisfaction with a given product).	The final score ultimately assigned to the entity is typically derived from a <u>single</u> respondent - a person who is expert in the factor being assessed – who often rates a variety of entities using the same standards of performance (e.g., a supervisor rates everyone in his or her work unit)

The advantage of using instruments to measure complex factors is that you are not dependent on the existence of other indirect indicators. By starting from scratch, you can identify whatever characteristics comprise that factor and construct individual items to assess them directly. The disadvantage of this approach is the extensive time and effort to develop, test, and administer the instrument to be sure it will produce accurate numerical scores. Here's the typical process involved;

- <u>Research and write prospective items</u>: The instrument development process generally requires the construction of 3 or 4 times the number of items that will appear in the final instrument and then using *item analysis* techniques to select the best items for use in the instrument. First search documents (e.g., other questionnaires or rating forms, reference texts, previous reports) and interview knowledgeable people (e.g., managers, performers, customers/users, technical experts, decision makers) for the kinds of information associated with the complex factor to be measured by the instrument. Then write the individual items using the format appropriate to the selected scale for each component of the complex factor. For example, if you were constructing an "employee satisfaction" questionnaire your research would probably identify components such as satisfaction with *management direction, resource support, working hours, facilities and materials,* *benefits*, etc. Each component would be represented by one or more written statements or questions in the item pool.

- <u>Select the best items using item analysis</u>: With an item pool containing about three or four times the number of items destined for the final instrument, the idea is to apply some type of systematic process that allows the developer to select the best items for the measurement purpose. This process, called *item analysis*, may involve either judges or tryout.

  - <u>Judges</u>: A group of 10 to 50 people are asked to rate the item statements in the pool against some qualitative characteristics (e.g., *relevance, importance, ease of understanding*) and those item statements having the best combination of characteristics are retained for the final instrument.

  - <u>Tryout</u>: The entire pool of items is formed into a large measurement instrument tried out by a group of typical respondents or knowledgeable raters. For example, an employee satisfaction questionnaire using all items in the pool are administered to a random sample of 30 employees. A total score from all the items is then derived to see which employees were most satisfied and which least satisfied. The results of each individual item are then analyzed to see how well they discriminated between those that scored highest on all items and those that scored lowest. The items that discriminated best are chosen for the final instrument.

- <u>Design the administrative technique</u>: Develop a strategy and prepare the necessary materials to assure that the instrument will be completed with honest and accurate responses.

      - <u>Questionnaires</u>: When administering questionnaires, you need to assure that a large proportion of the respondents complete the questionnaire. Since you're putting all this time and effort in the construction of the questionnaire to obtain accurate numerical scores, you don't want to fail because too small a percentage of respondents returned it to you completed. You must take special actions to attain a response rate in the range of 70 percent and above such as providing an incentive to respond, conveying the importance of the survey, making it easy to respond, and, as a last resort, conveying a sense that they will avoid a negative situation by responding. In short, do whatever it takes to avoid response bias by assuring that a large percentage of respondents complete and return the questionnaire.

  - Rating forms: There are three typical scenarios for administering rating forms:

    - A single rater directly observes the entities being assessed (one at a time) and immediately enters scores for each item on the rating form (e.g., a fire inspector rates the safety conditions of buildings).

    - A single rater observes the entities being assessed intermittently over an extended period of time and enters scores for each item on the rating form (e.g., annual performance appraisal by a supervisor).

    - Multiple raters directly observe the entities being assessed (one at a time) and immediately enter scores for each item on the rating form - the scores from the various raters are summed or averaged to determine the final score (e.g., assessment of figure skater performances or vendor selection).

    Under any of these scenarios, two requirements must be satisfied to assure the rating form will consistently yield accurate scores: (1) The *Rater* is very knowledgeable concerning the factors by which the entity will be assessed and has no preconceived notions or biases that will inappropriately influence the scoring process; and (2) *Standards of Quality* to which the entities will be compared are made easily accessible to the raters during the scoring process or have been internalized by them prior to the assessment by means of training and experience. In highly complex rating situations as used in assessing organizations for the Baldrige Award, it is not unusual for the raters to undergo extensive training and testing on applying the standards of quality before they are permitted to perform actual ratings.

- Test the final instrument for accuracy: To assure a questionnaire or rating form will produce accurate scores, it should be formally tested for *validity* and *reliability* before being used in an actual measurement situation.

  - Validity:

    Does the data collection method assign values to the unit of measure in a manner that truly represents the factor intended to be measured?

    Both questionnaires and rating forms can be tested for validity in a similar manner. Either (1) a group of people knowledgeable in the factor being measured can review the instrument and indicate their agreement that the score obtained from an actual use of the instrument will indeed represent the factor intended to be measured (i.e., *face validity*); or (2) it can be demonstrated that the items comprising the instrument are well aligned with documented references explaining the factor intended to be measured (i.e., *content validity*).

- Reliability:

  Would the scores produced by the data collection technique be consistent in spite of variations in irrelevant measurement conditions?

  Unlike validity, a test of reliability involves the actual administration of the instrument to a sample group of 30 or so test subjects. For a questionnaire, the major concern in reliability is the specific wording of the items comprising the questionnaire (e.g., might some seemingly inconsequential difference in language used in the writing of the items produce dramatically different scores?). The simplest way to test for this type of reliability is to use the *Alpha test*. This statistical test is based on the presumption that all items in the instrument are trying to measure the same thing (e.g., all items attempt to assess customer satisfaction in one way or another). It statistically checks for consistency in scores among all the items in the questionnaire. A statistically significant alpha statistic demonstrates that the particular wording of items does not seem to affect the final questionnaire scores. This means that the instrument produces consistent scores in spite of irrelevant differences in item wording. If the analysis shows poor reliability, then the items should be reviewed for such problems as ambiguous wording, double-barreled meanings, negatively worded items, as well as the use of loaded questions. The revised questionnaire needs to be tested for reliability again before it is formally administered.

  Although, testing a *rating form* for the alpha statistic is appropriate, a more critical source of inconsistency in the rating process involves the potential bias and knowledge of the rater. Rating scores should be independent of who is doing the rating. For example, the rating of product quality should reflect how well that product stacks up against its standards, not whether the rating is done by this product expert or that one. To test for this type of inconsistency among raters, rating forms are usually tested for *inter-rater reliability*. That is, two or more raters use the same rating form to assess a group of entities. The ratings scores are then compared statistically to assure consistency among the various raters. If the inter-rater reliability is high, then that rating process can be put into actual use with a single rater with confidence that the same scores would be obtained no matter who is doing the rating. On the other hand, if poor inter-rater reliability results from the test. The system has to be revised, possibly with more in depth documentation of the standards of quality or more intensive training or an improved rating form.

So, there you have it. You can measure the so-called unmeasurable if you're willing to put in the time and effort, and there is a critical need to do so. I'm not sure what the reason would be, but let's suppose you did have to put a number on the depth of the relationship between married couples. How could you do it? Would there be a single *indicator* (e.g., something like *percentage of discretionary time spent in each others' company*) that people would agree represents the depth the couple's relationship? Probably not, so maybe we could build an *index* score. We might construct a "Depth of Relationship Index" that would include indicators like *percentage of discretionary time spent in each others' company, frequency of physical touching (e.g., holding hands, sitting close), frequency of use of terms of endearment, length of time of the relationship, number of shared interests, outward signs of hostility, and number of supportive actions*. Of course we would have to convert each measure to a common scale and maybe assign weights to each indicator according to their relevance to a deep relationship. Could such an index score be considered a valid measure of the depth of a couple's relationship, and would it be practical to obtain scores on these indicators? If not, then we could always build an *instrument*, probably a questionnaire (since we're dealing with "feelings"), which would be administered to respondents to assess the depth of their relationship with their spouse. We would construct a pool of items on different aspects of one's feelings by researching books on interpersonal relationships and articles on marriage, and talking to marriage counselors along with people from a wide variety of relationships. The idea is to gain insights and to write as many items as practical on the various dimensions comprising the depth of a couple's relationship. With a large pool of items, we would then select the best ones, perhaps using a tryout in which sample respondents are administered a questionnaire comprising the entire pool of items. We would then use item analysis to select the individual items that best discriminated between respondents in intense relationships from those in the most tepid. Finally, we might test the questionnaire for (1) face validity (a panel of 20 counselors agree that the items do in fact represent the depth of marital relationships) and (2) use the alpha statistic to assure the questionnaire would result in reliable scores. If it doesn't pass these tests, we need to go back and modify the selection of items until validity and reliability can be demonstrated. At that time, we can say that we have a means to obtain a number that accurately represents the depth of someone's marital relationship. If we can do that, we can measure anything.

Jay Alden is a Professor of Systems Management at the Information Resources Management (IRM) College of National Defense University in Washington DC. He conducts courses on performance measurement and strategic management of websites. He is also the Chief Editor for the College's Community of Practice supporting federal Chief Information Officers. Dr. Alden previously was the Director of Executive Programs at the University of Maryland University College and the Director of Evaluation and Research at Xerox Corporation.