Tuning Up Your Rating Systems

John FitzGerald
www.ActualAnalysis.com
416-482-3603

Rating systems are used everywhere – to compare applicants for jobs, assess the job performance of people who have jobs, select students for university, estimate future capital expenditure, rate severity of handicaps, and so on. Properly constructed rating systems can in fact make decisions more dependable. However, many of the rating systems I run across are not constructed as effectively as they could be, and as a result they instead make decisions less dependable by providing misleading ratings.

By a rating system I mean a set of ratings of individual characteristics that are combined to produce a rating of a more general characteristic. For example, a typical rating system for appraising employee performance might combine separate ratings for quality of work, ability to meet deadlines, communication skills, interpersonal skills, and so on. These will then be added up to provide a single score that is a measure of employees' competence or usefulness to the organization. That sounds simple, but many problems can arise. Here I'll present some of them and show how you can deal with them.

1. The ratings can be too difficult or too easy

If we gave a class a test in mathematics and every student got a perfect score, we would have learned nothing about which students knew more about mathematics. Sometimes this result can be achieved quite innocently – an employer may want to find only his or her most highly promising employees However, any properly designed rating system will do that, and it will give you useful information about the other employees, too (for example, about how they could be encouraged to improve).

Similarly, a rating system on which every person, place, or thing rated gets a low score also gives you very little information with which you can distinguish the ability of suitability of those people, places, or things. You can evaluate the variation in scores on your rating system by putting them in a spreadsheet and using a couple of the spreadsheet's statistical functions. In Excel these functions are called SKEW and KURT (for kurtosis). Skew is, roughly speaking, a measure of the difference between the mean score and the median. If skew is greater than 1.00 OR less than –1.00 then the test is probably too difficult or too easy. Kurtosis is a measure of how much scores cluster around the mean score; if the kurtosis coefficient for your scores is greater than 1.00 then too many people are getting similar scores.

2. The value of the items can vary

I have seen rating systems where the highest possible score on one item will be 3, on another it will be 5, and on another it will be 10. If all three items are of the same difficulty then the score on the entire rating system will usually be determined by the item with the highest maximum score. That is, the system combines three items but only gets the effect of one. Furthermore, the effect of the one item will be reduced because the ratings on the other

items will add random amounts to the scores. The system will still pick out the highest and lowest performers, but distinguishing between the rest of the people or things rated will be much more difficult. This problem also occurs when weights are applied to items after they have been rated.

A quick way of assessing whether you have this problem is to put the ratings on each item in the system into a spreadsheet, add them up, and then use the Pearson correlation function (in Excel this is called PEARSON) to calculate the correlation coefficient for the relationship between each item and the total score. Ideally the coefficients will all be higher than .31. Items with a correlation lower than that are probably confusing the ratings. Items with a coefficient below zero are definitely confusing the ratings.

If you are using weighted ratings, you can calculate a new score based on unweighted ratings on each item and then correlate the unweighted rating of each item with the unweighted total score you will have discovered evidence of the problem discussed in the next section. If all the items are now adequately correlated you have solved your problem – as long as you use the unweighted scale from now on. If you still think some items are more important than others, you should set up different sets of items and evaluate them separately.

If the items are rated on different scales, try rating them on the identical scales and then checking he correlations.

If some unweighted or re-scaled items are still not correlated with the total score, you have found items that have the problem discussed in the next section.

3. The items can be unrelated to each other

A rating system is supposed to measure a characteristic or a trait – that is, a single characteristic or trait. Therefore, the ratings of the individual items should be similar to each other. If they're not, then combining them produces a meaningless rating.

Let's look at an extreme example. If for a number of cities you added together annual sales of shoes, the number of chairs in barber shops, and the floor space of furniture stores, you could claim to have a business index, but not many people would be interested in it unless you could demonstrate that those three variables were all signs of the same thing.

Sometimes items are unrelated because ratings on them don't vary adequately. Everybody may get the highest possible score on one of the rating, for example – if a rating doesn`t vary it can`t logically be correlated with any other measure. Then again, all your data may vary adequately, but measure several different things. This problem is often encountered in attitude or satisfaction surveys, but I have also seen it in databases which calculate other types of rating.

In attitude or satisfaction surveys, the problem is that responses to any single attitude item are influenced by many factors in addition to the attitude being assessed. Often the attitude will be less important than these other factors in determining the response to the question. Capital expenditure formulas, and other non-attitudinal ratings, can also have this problem.

Another difficulty with capital expenditure formulas is that they are intended to measure an abstract, and usually hypothetical, concept – need for capital investment. Often this will turn out not to be a single concept, but two or three.

You can assess internal consistency using the method described in the last section – putting the ratings on each item in the system into a spreadsheet, adding them up, then using the Pearson correlation function to calculate the correlation coefficient for the relationship between each item and the total score. Ideally the coefficients will all be higher than .31. If some aren't, decisions have to be made.

I recommend first of all that you remove all items whose coefficients are less than .20; if the coefficients are below zero you really must remove them, because they are seriously damaging your ratings. This is a do-it-yourself guide, but if you have people in your organization who can perform psychometric analysis, I recommend you turn them loose on your rating system – they'll perform some other helpful analyses, too. For example, they will be able to determine if you can keep all your items by arranging them into different sets. If you don't have people like that around, I will self-interestedly observe, as my one and only plug in this article, that one thing you could do is call or email me at the number or address given at the end of this newsletter.

4. What the system measures can be unstable

A well-designed rating system will still be of little use if what it's measuring is unstable. Instability can have a number of sources. In market research, for example, attitudes towards products will often be strongly affected by advertising campaigns. If we want to track changes in a rating system over time, we need to know that repeated ratings are similar. For example, you would not expect your IQ score to differ much over repeated testing. Test stability is another thing you assess with correlation coefficients. You need people or places or theoretical concepts to be rated at least twice at the same interval to assess stability. You simply calculate the Pearson correlation between the first and second set of scores. The coefficient should be at least .71; a coefficient of .90 or more is desirable.

5. Different raters can give different ratings

If the rating a person or object gets is dependent on who's doing the rating, the rating system will obviously be uninformative to some degree, often a great one. How we assess agreement between different raters will vary with the type of rating system, but I can assure you that *simple percentage agreement is not an adequate measure*, no matter how many people report it as if it is.

For example, people often report that agreement between a pair of raters was something like 90%. However, that figure needs to be compared with how often they would have agreed if they were using the rating system completely differently. For example, if two raters use a simple two-category system (pass/fail, for example), and they each put 90% of their ratings in one category and 10% in the other, then you'd expect them to agree by accident 82% of the time. Agreement of 90% doesn't look so impressive now, and a statistical test would be

needed to determine if it really is. For most types of rating system more powerful statistical analyses of agreement can be performed, too.

How did I get that figure of 82% in the last paragraph? If two raters each put 90% of their ratings in Category A, then the probability that they will agree just by accident in using this category is equal t 90% X 90% = 81%. Similarly, the probability that they will agree by accident in using Category B is 10% X 10% = 1%. You add 81% and 1% to get 82%.

Let's say that another rater actually put 80% of his responses in Category A and 20% in Category B, while a fourth rater put 70% of her responses in A and 30% in B. The calculations for these two raters would be:

a) 80% X 70% = 56%
b) 20% X 30% = 6%
c) 56% + 6% = 62%.

The percentage we have just calculated is the *expected agreement*. Now, if you expected agreement on 62% of the ratings and the two raters actually agreed on 64%, you would still suspect that they weren't really agreeing. To accurately decide whether actual agreement is better than chance you need to use a statistical test, such as the chi-square test. A less accurate guide can be obtained by assessing how much additional agreement was observed. If you expected 62% agreement and observed 64%, then agreement increased by only 2 percentage points out of a possible 38 (the difference between 100% and 62%). If actual agreement was 82%, then the difference would be greater than half the largest possible difference, and you'd feel more confident that the raters had agreed. The advantage of a statistical test is that it often can establish that lower percentages of agreement are signs of real agreement. Some tests can also consider ratings from several raters at once.

Oh, yes – you should probably have at least 100 pairs of ratings before you assess agreement. If you don't have enough real ratings to make, you can assess hypothetical cases.

6. The rating system doesn't measure what it's supposed to measure

This is of course the most serious drawback of any rating system, and unfortunately a common one. One way of evaluating a rating system is to correlate scores on it with scores on a known measure of whatever is being rated. A rating of sales agents' abilities should correlate with their sales, for example. There are two forms of this approach. In one you compare the rating to another measure taken at the same time (in psychometric terms, assessing concurrent validity) or with a measure taken later (predictive validity). You use the Pearson correlation again, and you assess the strength of the correlation by squaring it. Roughly speaking, a correlation of .50 improves prediction of sales by .50 X .50 = .25 = 25%.

You may also run across mentions of other types of validity. Content validity is simply the extent to which the items on a rating scale try to assess all the aspects of what is being rated. For example, a test of knowledge of nineteenth-century Canadian history may be reviewed to see that it contains items about all the historical events and analyses that the rater wants to

assess. However, content validity does not guarantee that the rating scale is measuring what it is intended to measure.

Construct validity is a more general version of concurrent validity. Face validity is simply a subjective assessment that the rating scale looks as if it would be a good measure, but it has no value in assessing the utility of the scale.

In short, getting the most out of rating systems is like getting the most out of anything else. We don't put bricks, mortar, concrete block, and wood into a pile and expect them to turn into a house without more attention from us, and adding up a group of ratings doesn't guarantee that we'll end up with a better rating, or any rating at all.

(more about me on the next page)

DON'T LET DATA MISINFORM YOU

People collect data so they can be informed, but often the data don't inform them at all. For example:

- I have repeatedly found in surveys that people said one thing made them happy about service while their other answers implied something else did.
- I have repeatedly found decisions being based on rating systems that don't rate accurately.
- I have repeatedly found people concluding that two groups had different degrees of satisfaction or different opinions when in fact there is little evidence that the groups do differ.
- And I have found much, much more.

But...*there is hope!* These problems can usually be solved by statistical analysis. Statistical skills are not widespread, But I've been exercising several of them daily for over 30 years, and I can use them to help you.

Trying to be informed by uninformative data is no fun. I can help you get rid of those fun-killing uninformative data.

Services:

- design of questionnaires and rating systems
- sampling and research design
- data analysis and reporting
- I am experienced with a wide range of evaluation topics conducted in co-operation with a wide range of groups: budget assignment, staff assignment, equity issues, drug and alcohol use, student recruitment, records management, computer use, opinion polling, foster care, evaluation of day care centres, selection procedures for special education, quality of working life, consumer satisfaction, rehabilitation etc.
- I am especially experienced in making rating systems more efficient. If you use rating systems to make decisions about either budget or staff, I can tell you whether you're collecting useful information and rating it properly.
- Extensive experience in assessing the adequacy of assessment procedures, including psychometric evaluation of placement instruments.
- Experienced in analysis of variance, factor analysis, and multiple linear regression. I never construct a regression equation by an automatic procedure, and I never use default criteria to extract a factor structure.
- Program logic modelling

John FitzGerald • 1170 Bay Street, #102 • Toronto, Ontario M5S 2B4
john@actualanalysis.com • 416-482-3603