Information

A User's Guide

No. 12 Fall 2011

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

How to Get Bad Ratings

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

By a rating scale 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.

The items can be unrelated to each other

A rating scale 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. The three measures are so obviously unrelated that their sum can't be a measure of anything.

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. 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-attitude scales, 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. The data used to assess this concept may therefore fall into two or three groups which are not necessarily related to each other.

A quick way of assessing whether your rating scale is internally consistent is to put the ratings on each item into a spreadsheet, add them up, 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 .3. If some aren't, decisions have to be made. If you have people in your organization who can perform psychometric analysis, turn them loose on your rating scale – they'll perform some other helpful analyses, too. If you don't, one thing you could do is call or email me at the number or address given at the end of this newsletter. Not only do I know how to do psychometric analysis but I can also explain it in plain English.

If a rating scale actually consists of two sets of items measuring two different things, it should be split into two rating scales, provided their internal consistency is adequate. Separate decisions should be based on each.

The scales can be too difficult or too easy

If we gave a class a test in mathematics and every student got a score of zero, 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 high score also gives you very little information with which you can distinguish the ability of suitability of those people, places, or things. To find differences between people, individual ratings have to vary, and you get the most variation with a rating scale of average difficulty.

The value of the items can vary

I have seen rating scales 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 scale will usually be determined by the item with the highest maximum score. That is, the scale 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 scale 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.

What the scale measures can be unstable:

A well-designed rating scale 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 scale 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.

Different raters can give different ratings

If the rating a person or object gets is dependent on who's doing the rating, the rating scale 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 scale, 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 raters was something like 90%. However, that figure needs to be compared with the percentage of agreement that would have been observed if they were using the rating scale completely differently. For example, if two raters use a simple two-category scale (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 scale more powerful statistical analyses of agreement can be performed, too.

The rating scale doesn't measure what it's supposed to measure

This is of course the most serious drawback of any rating scale, and unfortunately a common one. One way of evaluating a rating scale 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 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 scales 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.

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 scales 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 scales more efficient. If you use rating scales 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