Tips for Developing Evaluation Tools: Developing Items

Janet Senf, PhD

When you evaluate performance the goal is to get accurate information that would be the same no matter who was doing the rating. Thus, the instrument should be valid, i.e. it is measuring what you want to measure, and reliable, i.e. the results should be similar no matter who is doing the rating. In developing an evaluation tool, you must decide how you will use the information. Among other possibilities, evaluation tools can be used to record global impressions, to provide feedback, to improve performance, or to document accomplishment of skills. The tips that follow are intended to produce more valid and reliable tools to provide feedback for residents.

Evaluate one behavior at a time. When more than one behavior is evaluated there is no way to evaluate them differently.

> **Example:**
> The resident gathers accurate and essential information about patients. How would you rate a resident who gathers accurate information, but doesn't get all the essential information?
> **Better:**
> The resident gathers accurate information about patients. The resident obtains essential information about patients. The negative part of this process is that it creates a longer form. The positive aspect is that you can provide feedback that is specific.

Avoid global or judgmental terms as anchors. The more global the terms used to describe points on a scale the less accurate the rating will be unless there has been specific training about how to use the scale.

> **Example**: A scale that includes points identified as 1=Outstanding 2=Excellent 3=Good 4=Fair 5=Poor. Each of these terms means different things to different people.
> **Better:** A scale that describes the frequency of behaviors as 1=Never 2=Sometimes 3=Half the Time 4=Often 5=Always. Although there is still the possibility of variation based on individual interpretation, it should be less with this type of scale.

Limit the numbers in the rating scale. There should be enough numbers to reflect real variation in behavior, but not so many that the same rater looking at the same behavior might give a different rating.

> **Example**: A scale in which 1, 2, and 3 are unsatisfactory, 4, 5, and 6 are satisfactory and 7, 8, and 9 are superior. Because there are so many numbers and the differences within categories are not labeled, it becomes easier to use a different number within that category. A rater who previously gave a 1 or a 9 would probably do so again, however, a rating of 5 might be a 6 or an 8 might be a 7 on a second rating.
> **Better:** A five-point scale with each of the numbers labeled as described above.

Measure a single dimension. The anchors should reflect different points along the same continuum.

Example: A scale in which the anchors are 1=unsatisfactory 2=marginal 3=meets expectations 4=exceeds expectations and 5=exceptional. You can tell if it is the same continuum if the same words can be used in each category. In this case it would not make sense to say "unsatisfactory expectations."

Better: 1=Way below average 2=Somewhat below average 3=Average 4=Somewhat above average 5=Way above average. The reference point here is an average performance.

Balance the scale. The number of positive and negative responses should be the same.

Example: A scale in which 1=Outstanding 2=Superior 3=Very good 4=Good 5=Poor. If raters were selecting a category randomly, 80% of their selections would be favorable.

Better: 1=Outstanding 2=Good 3=Fair 4=Poor or better yet, the scale rating frequency of behaviors as described above.