2012

Developing Behavior-Based Rating Scales for Performance Assessments

Megan Paul
University of Nebraska-Lincoln, mpaul@unl.edu

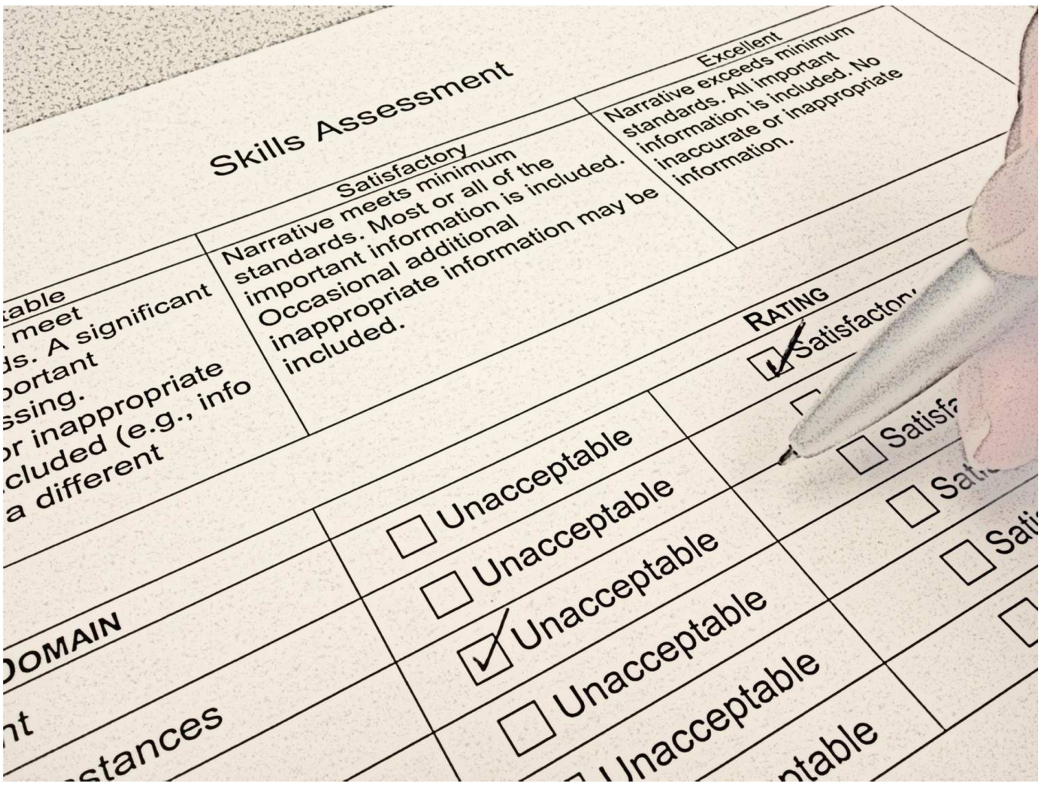
Michelle Graef
University of Nebraska-Lincoln, mgraef1@unl.edu

Kristin Saathoff
University of Nebraska-Lincoln

Developing Behavior-Based Rating Scales for Performance Assessments

Megan Paul, Michelle Graef, & Kristin Saathoff

National Human Services
Training Evaluation Symposium
May 2012



University of Nebraska Lincoln | CENTER ON CHILDREN, FAMILIES, AND THE LAW

Developing Behavior-Based Rating Scales for Performance Assessments

Performance Assessment: a subjective assessment of a process or a product, in either a simulated or real setting

A *performance assessment* is a subjective assessment of a process or a product, in either a simulated or real setting. Performance assessments are typically used as alternatives to either objective measures or to selected response measures (e.g., multiple-choice items). When there are no objective criteria for success, existing measures are inadequate, or a selected response measure isn't appropriate, performance assessments may be desirable. This booklet describes the process for developing performance assessments, with special attention to the development of behavior-based rating scales.

Determine the Purpose

The impetus for a performance assessment can come from several directions. Sometimes there is an interest in accomplishing some purpose (e.g., assessing training needs or evaluating the effectiveness of training), and then the next task is to determine what to assess to accomplish this purpose. Alternatively, there is often an interest in measuring a particular type of performance, with only a vague idea of the purpose and reason for doing so. Regardless of how things unfold, what is most important is that time and attention are dedicated to clearly identifying the purpose of the assessment. Here are some possible purposes for a performance assessment:

- Assess training or development needs
- Facilitate learning or improvement (i.e., use as a means of giving feedback)
- Evaluate training curriculum or delivery
- Assess the effect of training (i.e., gains in knowledge or skill)
- Evaluate implementation or effectiveness of a program (i.e., program evaluation)
- Ensure a certain level of proficiency has been achieved (e.g., certification)
- Distinguish among learners or performers (e.g., identify the top performers)

Seek Out SMEs

SMEs are *subject matter experts*: the people who know the subject matter best and can give you guidance, answer questions, and provide feedback throughout the development process. Consider them your best friends and always seek them out as a resource.

In a job training context, the best candidates are typically current or recent workers, supervisors, or administrators. Depending on the purpose, trainers and curriculum developers may also be appropriate. Although expertise is essential, it may not be sufficient. You may find that some SMEs are better suited to the task than others. Though it always helps to educate SMEs along the way, some excel in this area and others sometimes don't, due to lack of interest, time, or understanding of the process. Do your best to find the people that can contribute the most.

Subject Matter Experts: people who know the subject matter best and can give you guidance, answer questions, and provide feedback throughout the development process

Identify the Performance Target

The process of figuring out what to measure can vary widely. If you are lucky enough to have them, the results of a job analysis are the first best indicator of what performance is expected. If the assessment is intended to measure something taught in training, the curriculum should indicate the desired construct or performance dimensions. In either case, further clarification with trainers or other SMEs is sometimes necessary. In working with SMEs, you will find that they have anywhere from very broad to very specific targets in mind. Broad, and sometimes vague, targets include things like engagement, empowerment, cultural competence, facilitation, rapport building, documentation, communication, critical thinking, assessment, planning, and monitoring. As will be discussed in subsequent steps, getting to specific targets requires a deductive approach of translating general concepts into specific, observable criteria or behaviors. Alternatively, SMEs may have a series of more discrete criteria or behaviors in mind, and your goal will be to work backwards to figure out what the underlying categories or concepts are. At this point, all that is necessary is a more general understanding of what will be measured.

Decide Whether to Assess a Process, a Product, or Both

The process of identifying the performance target will probably reveal whether performance should be assessed through a process, a product, or both. For example, interviewing skills are probably best assessed by observing an actual interview, but court-report-writing skills are probably best assessed by reviewing a final court report. Some targets may require assessment of both a process and a product. For example, a case plan may be an important product to evaluate, but without evidence of the process, it may be hard to judge. What might otherwise look like an excellent case plan may have been created without a family's involvement, which is an inappropriate process. If the answer to this question isn't dictated by the performance target, consider which approach is more consistent with the intended purpose and which one is more practical, efficient, and feasible.

Plan the Assessment Task

Now is the time to think ahead about what type of assessment task you will use. At this point, the primary decision is whether the assessment task will be a structured exercise or a natural event. Because the primary purpose of the assessment task is to elicit the desired performance target, the decision should be based on which method will best accomplish this goal. Although it is important to make the task as realistic as possible, practical constraints or existing parameters may limit this. For example, if the purpose is to assess training needs of a new worker, it may be inappropriate to have the worker demonstrate a task on the job (e.g., by working with real clients or customers); instead a simulated exercise would be more appropriate. Alternatively, if the purpose is to give feedback to facilitate learning, and part of the training already includes an exercise in creating a specific product or demonstrating a process, the task is determined for you. In making this choice and in designing the details of the task, it is important to ensure that the assessment task elicits the desired process or product in a fairly reliable and standardized way. For example, if the performance target is conflict management, the situation must present conflict, probably of a certain quantity and type. More than likely, this could not be controlled in a natural environment, and a simulation would be necessary. Even for structured exercises, it is essential that all stimulus materials, conditions, prompts, and instructions elicit the performance of interest among all performers.

Select a Rating Scale

Knowing the assessment task and its parameters, you will want to think ahead about what type of rating scale might work best. Sometimes these decisions evolve as the details of performance become more apparent, but it is important to understand the options and keep them in mind as you go. The following four types of rating scales are described as behavior based, because of their focus on behavior. Despite the label, they can be used to rate product characteristics just as well.

Checklist. This scale includes a list of behavioral statements, and raters are asked to rate whether or not each behavior was exhibited. See Figure 1 for an example.

Figure 1:
Checklist Example

Use of Mechanical Restraints	No	Yes
Waist Belt		
• Positioned self behind youth	☐	☐
• Attached the waist belt snugly	☐	☐
• Wrapped any excess length around the waist belt itself	☐	☐
Handcuffs		
• Positioned in front of youth	☐	☐
• Applied handcuffs "double-bar-up" with key release points facing self	☐	☐
• Ran open handcuffs through steel loop on waist belt	☐	☐
Leg Irons		
• Positioned self to side of youth	☐	☐
• Applied leg irons "double-bar-up" with key release points facing self	☐	☐
• Properly sized leg irons – not too tight nor too loose	☐	☐

Behavioral Observation Scale (BOS). This scale includes a list of behavioral statements, and raters are asked to rate each behavior on a frequency scale (Latham, Fay, & Saari, 1979). See Figure 2 for an example.

Rapport Building	Never/ Rarely	Sometimes	Frequently/ Always
Uses slow, deliberate pacing/word speed			
Speaks in a soft voice (volume and tone)			
Uses clear and simple language			
Displays comfortable body posture/spacing			
Maintains eye contact/gaze (without staring)			

Figure 2:
BOS Example

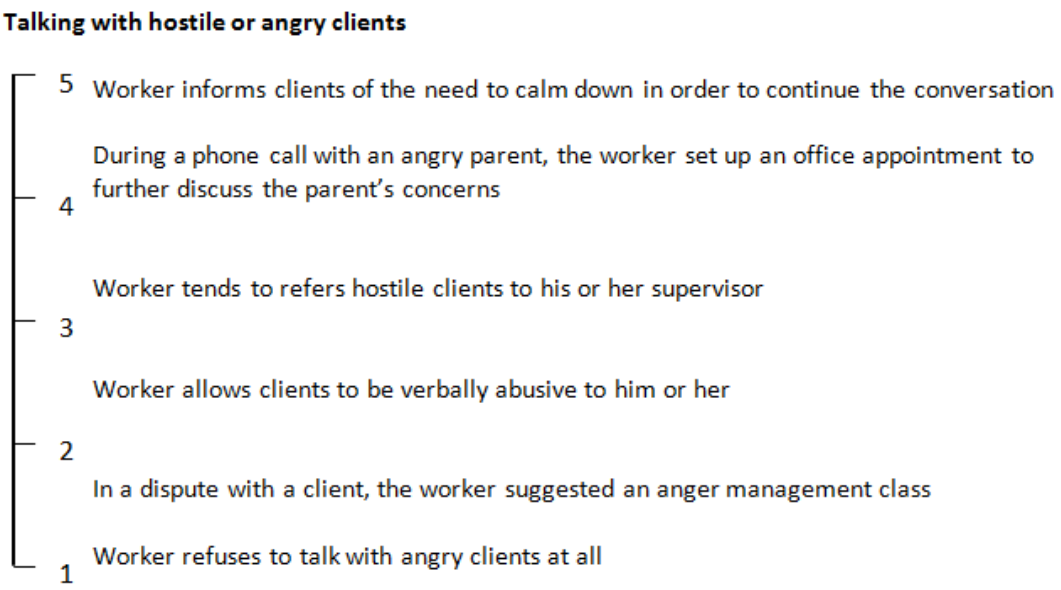
Behavioral Summary Scales (BSS). This scale includes a series of important performance dimensions, with general behavior descriptions anchoring different levels of performance effectiveness. Raters are asked to choose the rating that best describes an individual's performance (Borman, Hough, & Dunnette, 1976, cited in Borman, 1986). This is probably the format with which people are most familiar. See Figure 3 for an example.

Skill in asserting oneself			
➢Exhibits a hostile, blaming, accusatory, confrontational, or threatening response	➢Exhibits an aggressive response: Expresses own feelings, opinions, and needs in a harsh manner while devaluing or criticizing the feelings, opinions, and needs of others	➢Exhibits a passive response: Indirectly expresses own feelings, opinions, and needs in a vague or apologetic manner while deferring to the feelings, opinions, and needs of others	➢Exhibits an assertive response: Directly and appropriately expresses own feelings, opinions, and needs while respecting the feelings, opinions, and needs of others

Figure 3:
BSS Example

Behaviorally Anchored Rating Scales (BARS). This scale is similar to the BSS, except instead of general behavior descriptions, it includes specific behavioral exemplars (Smith & Kendall, 1963). Raters are asked to decide whether a given behavior they observed would lead them to expect behavior like that in the description (in fact, BARS were originally called *Behavioral Expectation Scales*). Thus, the observed behavior does not need to (nor would it be likely to) match the behavior descriptions in the scale. Because of the challenges with projecting expected behaviors based on observed behaviors, this approach is not recommended. See Figure 4 on the following page for an example.

Figure 4:
BARS Example



For guidance on choosing a rating scale, see Table 1 below. Keep in mind that you can use different types of scales in one assessment, depending on your needs.

Table 1:
Choosing a Rating Scale

If the behavior or characteristic will be exhibited…	Then you may want to create a…
once and is not likely to vary in quality	Checklist: Describe the behavior or characteristic and assess whether it was exhibited (*yes* or *no*)
once and is likely to vary in quality	BSS: Describe varying levels of performance quality and assess which level best describes the performance
multiple times	BOS: Include the quality in the behavior and then assess how frequently the desired behavior or characteristic was exhibited (e.g., *never* to *always*)

Detail the Performance Target

Now it is finally time to flesh out the details of the specific behaviors or product characteristics. Again, a job analysis or training curriculum will be informative, as will discussion with SMEs. The choice of rating scale will dictate what kind of behavioral descriptions to elicit from SMEs. For the most part, the only type of scale that requires extensive descriptions of all levels of performance is the BSS. A BOS will typically require only desirable behaviors, although if there are critical ineffective behaviors that need attention, they should be included as well. (Note, however, that the items

will have to be reverse-coded to ensure that frequent performance of a negative behavior results in a low score, whereas frequent performance of a positive behavior results in a high score.)

To help SMEs generate ideas, consider posing the following questions, as applicable:

- What behaviors or product characteristics separate good from poor performers?
- Think of a good/marginal/poor performer you know, or imagine the ideal/average/worst performer. What might he or she do? What would his or her products look like?
- Think of a time when a worker did a really good/mediocre/bad job. What did it look like?

The ideas generated by SMEs will need to be whittled down and shaped to arrive at specific anchors for the scale. Before doing this, you will need to decide what range of performance you want the scale to reflect. One consideration is the likely range of performance among those who will be assessed. How much variability in performance is anticipated? Within this range, what levels of performance are anticipated? For example, among novice performers, there might be a broad range of possible performance, with the average performance tending toward the middle or lower end. For more experienced performers, however, there might be a narrower range of anticipated performance, with the average performance tending toward the upper end.

The next consideration is what range of performance expectations you want to establish with the assessment; regardless of what behavior you anticipate seeing, what are the standards for performance? Be sure to avoid unreasonable expectations, especially those that go beyond what the job requires. In essence, you will want to consider these two questions: What *will* they do? What *should* they do (or not do)? Think about the answers in light of your purpose, and decide what range and levels you want to cover in the assessment. For example, a group of novices may rarely or never exhibit excellent performance, but if the purpose is to give feedback for improvement, the assessment should include anchors for excellent performance, even if they will almost never be used. Performers will then see what it takes to be an excellent performer and can strive to achieve it (or they will at least have a realistic impression of where they stand). Conversely, if the assessment is intended to ensure that a minimum performance standard has been met, the scale may not need to go beyond that minimum standard.

If you intend to use a BSS, you will need to decide on the number of rating categories before crafting all the anchors (of course, you will also need to do it with a BOS, but it can be deferred until later if you wish). Keeping in mind how the rating information will be used, you should determine how

many options will best capture meaningful differences in behavior. In most cases, more than five options is probably too many. Raters may not be able to make such fine distinctions, and having too many options causes the differences in ratings across performers to be more artificial than real. Conversely, it is possible to have too few options, which will artificially decrease or mask meaningful differences across performers. SMEs may be able to give some insight into what amount of discrimination is possible for the process or product in question. Aside from the standard rating categories, there may be some dimensions for which behaviors are so egregious that they need to be flagged for special attention. If this is the case, you may want to consider whether a *red flag* category might be useful as well.

If you are using a BSS, you may want to select shorthand labels for each category at this time (e.g., very poor, poor, marginal, good, very good). Note that the labels alone should not determine ratings; raters should be cautioned against relying on them to make judgments. That said, the labels need to be chosen carefully so as to prevent confusion and misinterpretation. When selecting labels, ensure that labels do not overlap and can be clearly distinguished. Also, if you have more than two categories, don't use labels that are technically dichotomous, such as unacceptable/ acceptable, unsatisfactory/satisfactory, or ineffective/effective.

At this point, you should be ready to refine the target performance. During this process, it is important to ensure that choices are driven by the intended purpose of the assessment and by specific job requirements. Without vigilance, it is possible to drift toward performance expectations that don't have much significance to actual job performance. Be sure to focus on frequent and important job activities or critical knowledge and skills. The following tips are intended to help guide the process:

General Tips

- Describe a performance continuum; ensure that the full range of performance is covered.
- Use clear and concrete language; avoid vague or ill-defined descriptions.
- Beware of oft-promoted action verbs (e.g., describe, define, discuss) that may not be the best indicators of the target performance.
- Use the same formula, format, and grammar across behaviors.
- Ensure that raters will have a clear and shared understanding of what each anchor means and that the anchors are distinct from one another.
- Avoid double negatives. For example, *never fails to make home visits.*

- Choose rating anchors that will best capture meaningful differences in the behavior or performance being evaluated. Especially when creating a BOS, it's easy to overlook the meaning of the different categories of frequency. For example, if there is no difference between something *never* happening and something *rarely* happening, you can make a single anchor, labeled *Never or Rarely.* However, if this is a meaningful difference that you want to know about, use each of them as a separate anchor.
- Consider the likelihood of each option being selected. If most of your responses are likely to be in the middle of your scale, such that the extremes are unlikely, you may need to expand the number of options (so as not to force everyone into one rating) or you may need to change the labels so they are not so extreme (e.g., *Frequently or Always*, instead of just *Always*).

BOS Tips

- Focus on single behaviors (or a collection of behaviors that co-occur). Avoid double- or triple-barreled descriptions that may deserve more than one rating.
- Ensure that it is logically possible for every option to be selected. Sometimes an option simply isn't viable and should be eliminated. For example, if you were to use a BOS to assess spelling, is it likely that a person would *never use proper spelling*?
- Don't include any frequency language in the behavior.

BSS Tips

- Use parallel language across performance levels.
- Identify the aspects that will vary across performance levels and stay focused on them. Don't shift focus by, for example, focusing on the frequency of a behavior in the "poor" category and focusing on the quality of a behavior in the "good" category. Pull the thread all the way across all levels of performance.
- Ensure that all performance has a place in the rating scale.
- If there are multiple behaviors or characteristics within a single category, make it clear to raters whether they are alternatives or requirements.
- To ensure consensus on which level of performance a behavior fits in, have a different group of SMEs rate each behavior on the intended scale (using labels only), and retain only those behaviors for which there is a minimum level of interrater agreement.

Once the rating scales are completed, there are several additional steps necessary to complete the performance assessment tool and process. You will need to fully develop the assessment task, determine the rating process, decide how performance will be scored and how the results will be used to achieve the purpose, select and train raters, and pilot the assessment before final implementation. For guidance on these issues, see the Recommended Readings section at the end of this booklet. Note also that if you developed a training assessment and discovered that the desired performance wasn't apparent from the curriculum, it's likely that the curriculum needs work. If it wasn't obvious to you, then it's probably not obvious to trainees either. The newly created performance expectations should be incorporated into training so that there is clear alignment between the training and the assessment.

References

Borman, W. C. (1986). Behavior-based rating scales. In R. Berk (Ed.), *Performance assessment: Methods and applications* (pp. 100–120). Baltimore, MD: The Johns Hopkins Press.

Latham, G. P., Fay, C. H., & Saari, L. M. (1979). The development of behavioral observation scales for appraising the performance of foreman. *Personnel Psychology, 32*, 299–311.

Smith, P. C., & Kendall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47(2)*, 149–155.

Recommended Readings

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Clauser, B. E. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement*, *24 (4),* 310–324.

Johnson, R. L., Penny, J. A., & Gordon, B. (2009). *Assessing performance: Designing, scoring, and validating performance tasks.* New York: Guilford Press.

Lane, S. & Stone, C. A. (2006). Performance Assessment. In R. Brennan (Ed.), *Educational Measurement*. (4th ed., pp. 387–431). Westport, CT: American Council on Education and Praeger Publishers.

Pulakos, E. D. (1991). Behavioral performance measures. In J. Jones, B. Steffy, & D. Bray (Eds.), *Applying Psychology in Business: The Handbook for Managers and Human Resource Professionals* (pp. 307 –313). New York: Lexington Books.

Stiggins, R. J. (1987). Design and development of performance assessments. *Instructional Topics in Educational Measurement, 6(3),* 33–42.