Available online at www.sciencedirect.com

ScienceDirect

Currents in Pharmacy Teaching and Learning 7 (2015) 656–661

Currents in Pharmacy Teaching & Learning

http://www.pharmacyteaching.com

Methodology Matters

# Measuring rater judgment within learning assessments—Part 1: Why the number of categories matters in a rating scale

Michael J. Peeters, PharmD, MEd, FCCP, BCPS*

*College of Pharmacy and Pharmaceutical Sciences, University of Toledo, Toledo, OH*

**Abstract**

Assessments will focus learning, and so it should be aligned with desired educational outcomes. Frequently human raters (content experts) are needed to judge many abilities of advanced learners. To use subjective human judgments effectively, both the number of categories in a rating scale and the design of rubrics are important. This first article of the Methodology Matters section discusses cognitive limits in rater judgments, rating scales, and their applications. The second part of this article considers use of a "mixed approach" to rubric creation; holistic and analytic rubrics are described, as is dual-processing theory to help explain the advocated mixed approach. After reading part 1 of this article, readers should be able to (a) discuss why a four-point rating scale is often preferred (using cognitive psychology, avoiding a middle category, and rating scale performance from Rasch Measurement), and (b) create a preferred rating scale application for a learning assessment in pharmacy education. After reading part 2 of this article, the readers should be able to: (a) recognize the differences between holistic and analytic rubrics, (b) discuss integration with cognition and dual-processing theory, and (c) create a rubric for a learning assessment using a mixed approach.



*Keywords:* Rating scale; Rater judgments; Pharmacy education; Learning assessment

## Situation

Frequently this author has been asked to suggest a rating scale for newly developed scoring instruments or has tried to prevent a proposed scoring scale from becoming too messy, complicated, or otherwise error-prone. A recent example was when creating a rubric within a committee for a professional organization. During this work and in relation to a prior report,[1] the committee had been tasked to assist colleges and schools of pharmacy with evaluating their faculty development in scholarly teaching. For this effort, a rating instrument was envisioned with items for aspects of scholarly teaching and was to be aligned with the recommendations from that prior committee report. Using this experience as an example, this article describes a process for creating valid and appropriate rating scales within rubrics that use rater judgments.

## Methodological literature review

Among many experts and generalists it is commonly accepted that, "assessments drive students' learning."[2–6] A practical application of this is the idea that many students are savvy and will learn what they need to succeed on a course's learning assessments.[2–8] For instance, students will study differently for multiple choice than for essay-based examinations.[2,9,10] Bloom's taxonomy,[11] which is routinely utilized when crafting educational objectives, is well-known to educators. Another simple and straightforward framework for learning assessments is Miller's pyramid of competence.[12] Figure 1 shows an adaptation of this pyramid for pharmacy education.[13] It transitions from recall-based multiple-choice examinations for basic understanding (knows),

* Correspondence to: Michael J. Peeters, PharmD, MEd, BCPS, FCCP, University of Toledo College of Pharmacy and Pharmaceutical Sciences, 3000 Arlington Ave, Mail Stop 1013, Toledo, OH 43614.

E-mail: michael.peeters@utoledo.edu

http://dx.doi.org/10.1016/j.cptl.2015.06.015

1877-1297/

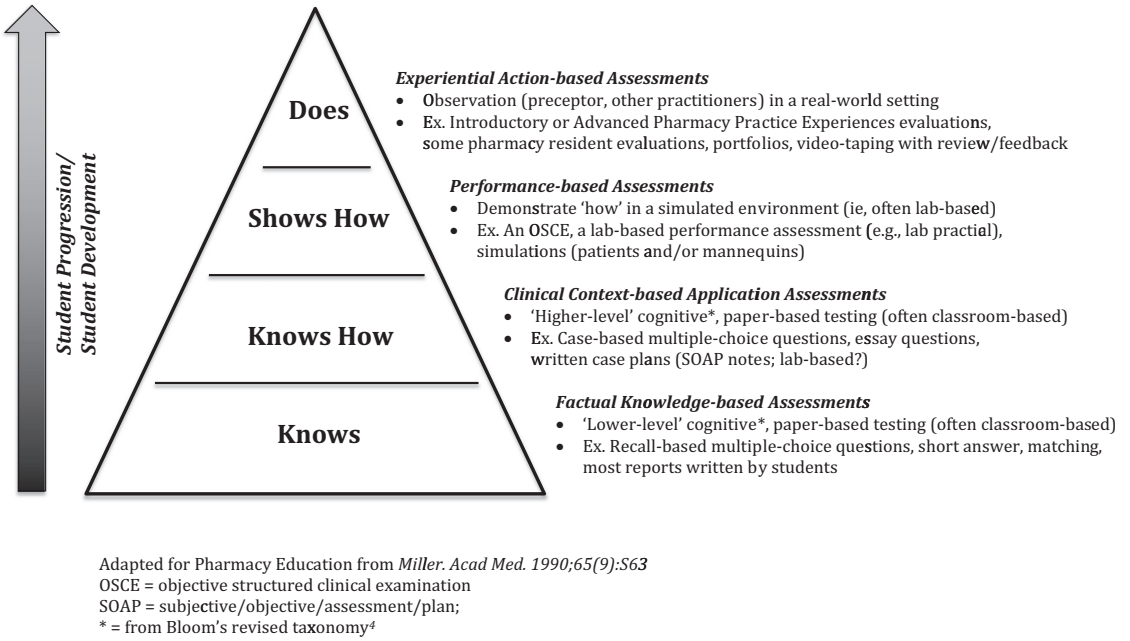


Fig. 1. Miller's pyramid of competence for learning assessments in pharmacy education. (Used with permission from Cor and Peeters.[13])

to more complex case-related examination questions (knows how) toward performance assessments in laboratory (shows how), and workplace (does) contexts. This pyramid is useful for understanding the types of assessments that should be used, based on the reason (or environment) for assessment. As one moves up Miller's pyramid, judgments by external raters play a larger and larger role in those assessment approaches.[2–4,8] As you might expect, obtaining raters' scoring effectively and reliably is fundamental in making valid conclusions from that learning assessment. It is paramount in this rater scoring process to be mindful that the number of categories in a rating scale will matter for any external rater-mediated learning assessment.

The unitary "validity" within the Standards for Educational and Psychological Testing is abbreviated for "construct validity,"[14] and foundational in rubric construction is the rating scale used to measure raters' judgments of learner performance. In this review, triangulation of evidence sources has been used.[15] The rest of this section details the intricacies associated with creating a rating scale from a number of perspectives and includes examples of literature supporting this recommendation.

## Cognitive psychology

Knowing more about raters' cognitive limitations can help us to create rating scales that will foster valid and reliable learning assessments. Almost a century ago, Symonds noted, "in psychological ratings it is useless to use a scale finer (with more categories) than the judge's ability to discriminate (among categories)."[16] Decades later, a landmark summary in psychology illustrated that people often cannot differentiate beyond seven categories, while for some tasks they manage even fewer categories (i.e., the "magic number" $7 \pm 2$).[17] Within this landmark article, "chunking" was coined to describe the concept of people's ability to intelligently group items in our limited short-term memory.[17–19] Thus, we should always use a rating scale with seven or fewer categories when humans must judge. More recently, this number was downsized to $4 \pm 1$ as a standard by which to operate.[18,19] Given that there is an inherent limitation on short-term memory of human raters, using more categories only asks each rater to subconsciously "chunk" the larger scale into $4 \pm 1$ categories of his/her own intelligent choosing, which does not necessarily result in the same item-groupings across all raters.

The following serves as an example of "chunking" in practice and its unintended effects: a ten-point scale was used in a performance-based learning assessment to score "oral communication." Raters likely chunked that ten-point scale for oral communication into roughly four portions. A rater may have categorized it as 1–3, 4–6, 7–8, and 9–10. This rater judged a performance as "borderline" and assigned a score of six (a six is in the second chunk for this rater). Meanwhile, a second rater may have used 1–3, 4–5, 6–7, 8–10, and also scored a six; though meant this score as a better "acceptable" performance (a six is in the third chunk for this rater). It is troublesome that performances that are judged to be qualitatively different by the raters were scored the same using the ten-point scale. In this