LETTERS

Summative Evaluations When Using an Objective Structured Teaching Exercise

Sturpe and Schaivone's primer on objective structured teaching exercises (OSTEs) was a timely addition to the pharmacy education literature.¹ The article cogently pointed out notable needs for effective improvements in faculty development and many "how-to" OSTE elements for pedagogical faculty development. Building off these ideas, we would like to add to the conversation by expanding on reliability needs (ie, consistency and fairness) with this type of assessment.

The OSTE is an elegant extension of the objective structured clinical examination (OSCE) technique. Such examinations are typically used to summatively assess pharmacy students' clinical abilities. In an OSCE's high-stakes context, achieving high levels of reliability is imperative. Generalizability theory (G-theory) is a gold-standard means to quantify reliability with this type of testing. Generalizability theory provides a framework to tease apart variation resulting from assessment aspects, such as raters, scoring instrument components, and each specific case context that contribute to total score variability.^{2,3} This theory demonstrates how context specificity leads to variation in performances based solely on differences in how students experience or are treated from one context to the next (ie, different raters and/or station scenarios). In recent decades, notable developments describing and examining context specificity within assessments have occurred.⁴⁻⁶

For example, G-theory analyses with data from OSCEs and OSTEs show that increasing the number of stations and/or examiners in a scoring scheme reduces score variation attributable to these design elements and subsequently improves reliability substantially.^{2,5-7} Taken together, these findings suggest that if colleges and schools of pharmacy move toward using OSTEs for summative purposes, OSTE designers must pay careful attention to the number of stations and raters used to produce overall OSTE scores. Thus, pharmacy education should be moving away from single-rater/single-station models of performance assessment towards models with more stations and more raters to improve reliability.

We commend Sturpe and Schaivone for discussing how an OSTE can be used for formative assessment (ie, ongoing feedback and faculty development) and summative assessment (ie, faculty/preceptor evaluation). In a setting of formative faculty development, feedback is more important than high-level reliability.⁸ Sturpe and Schaivone

eloquently describe this formative development goal. However, if an OSTE were used in evaluation or as an outcome in research, high-level reliability and avoiding measurement error would become imperative.³ Ultimately, we emphasize that, as Sturpe has noted elsewhere with summative OSCEs, more stations should be used to achieve acceptably-high reliability.⁹ By accurately viewing OSTEs as a version of OSCEs, the same principle applies. If an OSTE is to be used for summative evaluation, multiple stations and raters are needed (with possibly more than 3-5 stations;⁷ in general with fewer stations, more raters are needed—so numerous raters if only 3 stations are used, knowing that for reliability more stations is often much better than more raters in each station¹⁰).

Michael J. Peeters, PharmD, MEd^a

Conor P. Kelly, BSPS^a

M. Kenneth Cor, PhD^b

^a University of Toledo College of Pharmacy and Pharmaceutical Sciences, Toledo Ohio

^b University of Alberta Faculty of Pharmacy and Pharmaceutical Sciences, Edmonton, Alberta, Canada

REFERENCES

1. Sturpe DA, Schaivone KA. A primer for objective structure teaching exercises. *Am J Pharm Educ.* 2014;78(5):Article 104.
2. Norman G, Eva KW. Quantitative research methods in medical education. In: Swanwick T, ed. *Understanding Medical Education: Evidence, Theory and Practice.* 2nd ed. Chichester, UK: Wiley Blackwell; 2014: 349-369.
3. Peeters MJ, Beltyukova SA, Martin BA. Educational testing and validity of conclusions in the scholarship of teaching and learning. *Am J Pharm Educ.* 2013;77(9):Article 186.
4. van der Vlueten CP. When I say . . . context specificity. *Med Educ.* 2014;48(3):234-235.
5. Eva KW. On the generality of specificity. *Med Educ.* 2003;37(7):587-588.
6. Peeters MJ, Serres ML, Gundrum TE. Improving reliability of a residency interview process. *Am J Pharm Educ.* 2013;77(8): Article 168.
7. Quick M, Mazor K, Haley HL, et al. Reliability and validity of checklists and global ratings by standardized students, trained raters, and faculty raters in an objective structured teaching exercise. *Teach Learn Med.* 2005;17(3):202-209.
8. Cox CD, Peeters MJ, Stanford BL, Seifert CF. Pilot of peer assessment within experiential teaching and learning. *Curr Pharm Teach Learn.* 2013;5(4):311-320.
9. Sturpe DA. Objective Structured Clinical Examination in doctor of pharmacy programs in the United States. *Am J Pharm Educ.* 2010;74(8):Article 148.
10. van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardized patients: state of the art. *Teach Learn Med.* 1990; 2(2):58-76.