Available online at www.sciencedirect.com

ScienceDirect

Currents in Pharmacy Teaching and Learning 6 (2014) 298–303

Currents in Pharmacy Teaching & Learning

http://www.pharmacyteaching.com

Short communication

Inter-rater reliability and false confidence in precision: Using standard error of measurement within PharmD admissions essay rubric development

Michael J. Peeters, PharmD, MEd, BCPS[a,*], Kimberly A. Schmude, PharmD[a], Caren L. Steinmiller, PhD[a,b]

[a] University of Toledo College of Pharmacy & Pharmaceutical Sciences, University of Toledo, Toledo, OH
[b] Department of Psychiatry & Behavioral Neurosciences, Wayne State University School of Medicine, Detroit, MI

Abstract

The Accreditation Council for Pharmacy Education requires that written communication be assessed in Doctor of Pharmacy admissions processes. Reliability is a standard for ethical testing, and inter-rater reliability with scoring essays necessitates continued quality assurance. Both inter-rater consistency and inter-rater agreement are part of inter-rater reliability and so both need scrutiny. Within our admission process, we analyzed inter-rater reliability for faculty rater essay scores from 2008–2012 using intraclass correlation (ICC) for consistency and standard error of measurement (SEM) for agreement. Trends in these scores were examined to evaluate the impact of rubric implementation, revisions, and rater training integrated over the course of those five admission cycles. For regular admission (RA) candidates, an analytic rubric was implemented in 2009. Scoring without a rubric began with an ICC of 0.595 (2008) and improved to 0.860 (2012) after rubric implementation, revisions, and rater training. In a separate but similar process for contingent admission (CA) candidates, a holistic rubric was implemented in 2010. The ICC for CA essay scoring before rubric was 0.586 (2009), and it improved to 0.772 (2012). With both rubrics, inter-rater agreement (using SEM) improved with smaller scoring scales (i.e., 4-point > 20-point > 50-point). In our experience, rubric implementation and training appeared to improve inter-rater consistency, though inter-rater agreement was not improved with every rubric revision. Our holistic rubrics' 4-point scale was most precise for both inter-rater consistency and inter-rater agreement. Our rubrics with larger scoring scales appeared to foster false confidence in precision of scores—with larger variation in scores introducing more measurement error.

Keywords: Psychometrics; Reliability; Rubrics; Pharmacy education; Admissions

Introduction

It is very important for pharmacists to communicate effectively; therefore evaluating written communication is a required component in the Accreditation Council for Pharmacy Education's (ACPE's) standards for Doctor of Pharmacy (PharmD) programs.[1] For any evaluation, validity is a central issue and reliability is a substantial concern for validity.[2] To improve reliability with subjective essay grading, developing a rubric can be helpful and is suggested.[3] Rubrics may follow two general permutations —holistic or analytic rubric types. Requiring addition of rubric domains, the scoring scale for an analytic rubric can be complex and has numerous scale points such as 10, 20, or 50 points for the entire essay. Meanwhile, a holistic rubric often uses a less complex, smaller scoring scale (such as 4, 5, or 6 points) for the entire essay. While either rubric type can demonstrate sound reliability and validity,[4] not every rubric will inherently be reliable and helpful.

The data from this manuscript was previously presented as a poster at the 2012 AACP Annual Meeting.

* Corresponding author: Michael J. Peeters, PharmD, MEd, BCPS, University of Toledo College of Pharmacy & Pharmaceutical Sciences, 3000 Arlington Ave, MS 1013, Toledo, OH.
E-mail: michael.peeters@utoledo.edu

1877-1297/14/$ – see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cptl.2013.11.014

Moreover, both training and minor rubric revisions based on rater feedback may improve rubric reliability yet further,[3] but like initial rubric development, success is not guaranteed regardless of how much training or revision occurs. Ongoing continuous quality assurance for training and revision is needed to ensure success.

Within health profession program admissions, the Medical College Admission Test (MCAT) has used a written response section for the past couple of decades. With prior pilot testing, a central issue was the need for reliability in scoring.[5] Notably, holistic scoring was used [as does the current Pharmacy College Admission Test (PCAT)].[6,7] Unfortunately, aside from Mitchell's MCAT article just described, Salvatori's review of admission tools for health professions specifically addresses written submissions and states, "there is a paucity of literature on the use of essays."[8] This paucity may have resulted, in part, from dissimilarities in how the essays are reviewed and scored locally at each institution or by a single external testing organization (e.g., MCAT and PCAT). Therefore understanding an important yet subtle-sounding difference between inter-rater consistency and inter-rater agreement seems imperative.

Inter-rater reliability has relative and absolute indices.[9,10] Relative indices reflect inter-rater consistency, with intraclass correlation (ICC) or Cohen's kappa most commonly reported. Less often reported are absolute indices such as standard error of measurement (SEM); these may also be referred to as inter-rater agreement indices. For reliability of an educational assessment, SEM is suggested[11–13]:

$$\text{SEM} = \text{SD} \times \sqrt{(1-\text{reliability})}$$

where, SD is the standard deviation of essay scores among raters; reliability is by ICC in this evaluation.

Once familiar with it, this coefficient should seem intuitive and helpful to educational assessment; SEM is expressed in the same numeric values (or raw-score points) as the assessment's point scale, describes the confidence that decision-makers should have in a specific score of an assessment, and is a similar concept to the 95% confidence interval used in other medical literature. Recall using a parametric distribution with ± 1 standard deviation (SD) with 68% of data and ±2 SD with 95% of data. For example, a 16 on a test with a mean of 14 ± 3 SD and ICC of 0.8 (and so calculated SEM = 1.3) would suggest 68% confidence that this 16 is greater than the mean of 14 ± 1 SEM (range: 12.7–15.3); however, we are not 95% confident that the score is higher than the mean (i.e., 95% of 14 ± 2 SEM; range includes 16). The smaller an SEM is, the more confident we are in the precision of that scoring.[12] With SEM, we can be alerted to error (or imprecision) in a test score compared with others close by—such as a score of 15 and scores of 13 or 16; is there any meaningful difference beyond error among these scores?

Recognizing the importance of these psychometric principles, we sought to apply them to our admission process. Herein, we share our experience in that application. During our admissions process, three faculty raters evaluated written communication for each applicant's essay. Within two parallel processes for admission (defined further in the methods section), separate subcommittees of our Admissions Committee created separate rubrics for essay scoring. An analytic rubric was developed for regular admission (RA) applicants (i.e., at least college sophomores), and a holistic rubric was developed for contingent admission (CA) applicants (i.e., high school seniors). After each rubric was implemented and used, we made small revisions to those rubrics using post-use feedback from raters. In addition, training for raters was initiated. In this evaluation, we quantified inter-rater reliability of essay scoring while implementing two different rubric types within our admission process. This had two dimensions—inter-rater consistency and inter-rater agreement. We then evaluated the trends of inter-rater reliability with our rater training and rubric revisions based on rater feedback.

Methods

Design

This study was approved by our Institutional Review Board and evaluated reliability of scoring for admission essays during the 2008–2012 admission cycles. In this study, different rubric developments transpired in the two separate PharmD program applicant groups. College-level regular admission (RA) applicants, at a date and time designated by the College, wrote a 50-minute essay response to a question in a proctored classroom on campus. That essay prompt delved into their prior experience with an aspect of professionalism, and the specific question changed most years. These essays were usually of 250–300 words. Meanwhile current high school contingent admission (CA) applicants, if accepted, could be conditionally admitted to our PharmD program directly from high school, with their continued eligibility contingent upon maintaining a 3.5 GPA in their pre-PharmD university coursework. These applicants completed an online application that included a single-access, 120-minute timed essay of 250–500 words (with most being ~300 words). This CA essay prompt was similar for the past few years, though was not the same prompt given to RA applicants.

Instruments

Along with the parallel essay writing by the two applicant groups, two different rubrics were created for evaluating admission essays. Separate subcommittees of the Admissions Committee convened to develop the two different rubrics, and both were aligned with ACPE's admission requirement for evaluating written communication. The subcommittee members' experience with scoring these essays ranged from two to five years, with an average of three years, and included a member with advanced training