

BRIEF

Inter-rater Reliability of a Clinical Documentation Rubric Within Pharmacotherapy Problem-Based Learning Courses

Kristin R. Villa, PhD, PharmD, MS,^{a,b} Tracy L. Sprunger, PharmD,^c Alison M. Walton, PharmD,^c Tracy J. Costello, PharmD,^{c,d} Alex N. Isaacs, PharmD^b

^a University of Kansas, School of Pharmacy, Lawrence, Kansas

^b Purdue University, College of Pharmacy, West Lafayette, Indiana

^c Butler University, College of Pharmacy & Health Sciences, Indianapolis, Indiana

^d Family Medicine, Community Health Network

Submitted April 8, 2019; accepted December 30, 2019; published July 2020.

Objective. To evaluate a clinical documentation rubric for pharmacotherapy problem-based learning (PBL) courses using inter-rater reliability (IRR) among different evaluators.

Methods. A rubric was adapted for use in grading student pharmacists' clinical documentation in pharmacotherapy PBL courses. Multiple faculty evaluators used the rubric to assess student pharmacists' clinical documentation. The mean rubric score given by the evaluators and the standard deviation were calculated. Intra-class correlation coefficients (ICC) were calculated to determine the inter-rater reliability (IRR) of the rubric.

Results. Three hundred seventeen clinical documentation submissions were scored twice by multiple evaluators using the rubric. The mean initial evaluation score was 9.1 (SD=0.9) and the mean second evaluation score was 9.1 (SD=0.9), with no significant difference found between the two. The overall ICC was 0.7 across multiple graders, indicating good IRR.

Conclusion. The clinical documentation rubric demonstrated overall good IRR between multiple evaluators when used in pharmacotherapy PBL courses. The rubric will undergo additional evaluation and continuous quality improvement to ensure that student pharmacists are provided with the formative feedback they need.

Keywords: evaluation, rubric, inter-rater reliability, clinical documentation

INTRODUCTION

Practicing pharmacists need to accurately and holistically develop evidence-based recommendations as well as document their findings in order to contribute to the delivery of patient care.^{1,2} Colleges of pharmacy are incorporating training in clinical documentation to ensure that learners are following the Pharmacists' Patient Care Process (PPCP) in the delivery of care, as documentation and communication are two core principles of the PPCP.^{2,3} To properly incorporate clinical documentation, colleges of pharmacy use evaluation as part of their teaching and learning processes in addition to curriculum design; however, this can be a challenging task.³⁻¹¹

One challenge to evaluation globally, but specifically with clinical documentation, is the subjective nature of grading and the variability in grading between multiple

evaluators.⁴⁻¹¹ Leniency and stringency in evaluation has been described as a "hawk-dove effect," where evaluators tend to favor either lower or higher scores.^{4,5} Rubrics provide a standardized method for assessment and are often used as evaluation tools for student performance.⁷⁻¹¹ Rubrics have been evaluated within diverse pharmacy education settings, including skills laboratories, didactic courses, and experiential education.¹²⁻²¹ The literature supports the use of rubrics to enhance inter- and intra-rater reliability in evaluating student performance.¹²⁻²¹ One study found that a majority of colleges of pharmacy used holistic and analytic rubrics for assessing clinical documentation; however, the reliability of these tools was not evaluated.¹²

Rubrics should meet educational standards for reliability, which can include inter-rater reliability (IRR).⁷ Using rubrics that meet reliability standards supports consistent evaluations, better judgment of performances, and improved self-assessment.⁷⁻¹¹ Open-ended assessments with multiple correct responses, which is frequent

Corresponding Author: Alex N. Isaacs, Purdue University College of Pharmacy, 640 Eskenazi Ave, Indianapolis, IN 46202. Tel: 317-880-5423. Email: isaacs5@purdue.edu

with clinical scenarios, often leads to lower calculated reliability for rubrics.⁷ However, when a rubric is used for formative assessments rather than summative assessments, lower levels of reliability are acceptable.⁷

With the use of open-ended formative assessments and multiple evaluators, continual evaluation and quality improvement of rubrics within a pharmacy curriculum are essential.^{3,14,22} The goal of this study was to the IRR of a clinical documentation rubric used by multiple evaluators in pharmacotherapy problem-based learning (PBL) courses at one institution.

METHODS

At Butler University College of Pharmacy & Health Sciences, case studies courses are required, sequential PBL companion courses to the pharmacotherapy lecture courses for second (P2) and third professional (P3) year student pharmacists. Each case studies course had one professional class of student pharmacists divided into eight sections, and subsequently subdivided into groups of five or six student pharmacists. Each section was facilitated by one faculty member and one pharmacy resident. During weekly three-hour sessions, student pharmacists worked together to develop a pharmacotherapeutic plan for a patient case from a published casebook.²³ Each weekly class session included a unique patient case on a different disease state, mirroring content in the companion pharmacotherapy lecture courses. After completion of each class session, student pharmacists had 24 hours to submit individual clinical documentation in the form of a care plan or SOAP (subjective, objective, assessment, and plan) note.

Prior to using the rubric, student pharmacists' clinical documentation was evaluated by one faculty member or resident using a pass/fail scale and without clearly defined performance criteria, resulting in grading variability. Feedback from student pharmacists and evaluators prompted our school to begin the quest to standardize evaluation. Permission to modify an existing instrument was obtained from the University of North Carolina (UNC) Eschelman School of Pharmacy.²⁴ When initially incorporated at Butler three years prior to the current study, the rubric was modified by course faculty member to mirror existing terminology as well as to adjust the distribution of scoring. No additional modifications were made between the time the rubric was initially introduced and when the study began.

The modified rubric used for the study contained nine content areas for clinical documentation (Table 1). Each content area was divided into four competency levels, which were assigned by the evaluator based on defined performance criteria. Evaluators were then asked

to score student pharmacists' clinical documentation submissions using the revised rubric. Based on their performance, student pharmacists received an overall score between zero and 10. Each week, the submissions were evaluated by one of the course section faculty members or residents. Ideally, within one week after submission and prior to the next class, the completed rubric containing quantitative and qualitative feedback was provided to the student.

The rubric was used as the primary evaluation method in these courses prior to and during the study. Each semester, faculty members and residents were trained in how to evaluate student pharmacists' performance during a two-hour course introduction, which included 30 minutes dedicated to discussing the rubric. During the training session, the faculty members assessed a sample assignment using the rubric to model and standardize scoring between evaluators. While a copy of the rubric was given to student pharmacists prior to the course and feedback was subsequently provided to them after each submission was graded, no additional training was provided to students on how to use and learn from the tool.

Data were collected from pharmacotherapy PBL courses taken by P2 and P3 student pharmacists during one academic semester. Deviating from the traditional approach of one evaluator for each submission, both the faculty member and resident from each section were requested to evaluate submissions for three unique cases from each course. Only completed rubrics with evaluations from two evaluators were included in the analysis; rubrics that contained a scoring error or were missing two or more evaluator scores were excluded. Evaluators included first- and second-year pharmacy residents as well as diverse pharmacy practice faculty members at the assistant and associate level. Three assignments per professional year were initially selected to have two facilitators evaluate each submission; however, one assignment in the P3 year was excluded because of missing evaluations. Clinical topics for P2 student pharmacists included contraception, alcohol withdrawal, and diabetes, while clinical topics for P3 student pharmacists included community-acquired pneumonia and sexually transmitted infections.

Descriptive statistics were calculated using SPSS Statistics, version 25 (IBM Corp; Armonk, NY). Paired sample *t* tests were used to compare mean scores for each data set. A *p* value of less than .05 was considered significant. The intra-class correlation (ICC) coefficient was calculated to determine the IRR of the rubric overall and of its subcomponents. The ICC calculations were made using a one-way random effect model and single measures mean calculation (ICC [1,1]) as this provided the

Table 1. Detailed Clinical Documentation Rubric Used to Evaluate the Performance of Second- and Third-Year Doctor of Pharmacy in Courses and Study

Area (% of overall score)	Not Acceptable (0%) Comments Required	Needs Improvement (70%) Comments Required	Competent (85%)	Excellent (100%)
CrCl (5%)	Incorrect	-	-	Correct
Subjective/ Objective (5%)	Mostly incomplete, inaccurate or missing.	Poorly organized and/or limited summary of relevant information. Evidence listed, yet not differentiated.	Well organized; partial but accurate summary of relevant information. Major points included.	Complete and concise summary of evidence. No extraneous information present. Patient specific.
Etiology & Contributing Factors (5%)	Mostly incomplete, inaccurate or missing information.	Poorly organized and/or limited summary of relevant information (includes extraneous information).	Well organized; accurate summary of relevant information. Minimal extraneous information.	Complete and concise summary of relevant information. No extraneous information present. Patient specific.
Assess Need for Therapy & Current Therapy (10%)	Mostly incomplete, inaccurate or missing. Current therapy is not evaluated.	Poorly organized and/or limited summary of pertinent information. Disease state staging inaccurate (if applicable).	Well organized; partial but accurate summary of pertinent information. Accurate assessment of disease staging and progression if treatment not initiated/adjusted. Accurate assessment of current therapy.	Complete and concise summary of pertinent information. Need for therapy patient specific and appropriate discussion of disease progression if not treated.
Evaluation of Treatment Options (15%)	Evaluation of treatment options missing, minimal, or inaccurate.	Not all appropriate agents evaluated. Includes some inaccurate or extraneous information.	Appropriate agents chosen to evaluate. Includes minimal extraneous information. Evaluation accurate and risks/benefits weighed.	Complete and concise evaluation of treatment options. Patient specific. No extraneous information. Primary literature or guidelines referenced.
Treatment Goals (10%)	Mostly incomplete, inaccurate goals, or missing.	Some of the appropriate therapeutic goals listed.	Most of the appropriate therapeutic goals listed.	Complete and appropriate therapeutic goals.
Treatment Plan ^a (25%)	Mostly incomplete, inaccurate, or missing.	Partially complete recommendation for treatment.	Changes to treatment plan mostly appropriate and documented properly.	Complete and appropriate therapeutic plan recommendation including recommendation for current therapy.
Follow-up & Monitoring (10%)	Mostly incomplete, inaccurate, or missing.	Some appropriate monitoring. Inclusion of some unreasonable monitoring frequencies.	Most of the appropriate monitoring parameters and reasonable frequencies listed.	Complete and appropriate monitoring parameters and frequency listed.

(Continued)

Table 1. (Continued)

Area (% of overall score)	Not Acceptable (0%) Comments Required	Needs Improvement (70%) Comments Required		
		Competent (85%)	Excellent (100%)	
Patient Education (15%)	Inappropriate counseling information or not addressed.	Patient education addressed, but missing information presented not prioritized, patient specific, or patient friendly terminology.	Most of the important counseling points addressed. Patient specific. Extraneous information limited. Patient friendly language used.	Complete, well-organized, patient friendly language used. Important drug information, side effect information provided appropriately.

Abbreviations: CrCl=creatinine clearance

^a Treatment plans which result in harm to the patient will result in an automatic zero for the assignment

most conservative ICC estimates.²⁵ Scores were calculated for individual cases, by professional year, and collectively to assess the rubric IRR. Higher ICC values indicated stronger IRR, which was suggestive of stronger agreement between evaluators. The ICC scores less than 0.4 indicated poor IRR, while scores between 0.4 and 0.6 indicated fair IRR, scores between 0.6 and 0.7 indicated good IRR, and scores between 0.8 and 1 indicated excellent IRR.²⁶

RESULTS

Of the 789 clinical documentation submissions by P2 and P3 student pharmacists, completed rubrics from two evaluators were available for 317 (40%), which were included in the analysis. Of these, 234 submissions (74%) were completed by P2 student pharmacists and 83 submissions (26%) were completed by P3 student pharmacists. There were no significant differences in mean initial evaluation scores for the sample (mean=9.1, SD=0.9) compared to the mean second evaluation scores for the sample (mean=9.1, SD=0.9).

The overall ICC for all five assignments was 0.7 ($p < .01$; 95% CI, 0.6-0.8), indicating good IRR. The ICC for evaluations completed by P2 and P3 student pharmacists using the rubric were 0.7 ($p < .01$; 95% CI, 0.7-0.8) and 0.5 ($p < .01$; 95% CI, 0.4-0.7), indicating good and fair IRR, respectively. Mean and standard deviation scores by class and topic, as well as each individual ICC, are presented in Table 2. No significant differences were found in evaluators' mean scores for any individual assignment.

DISCUSSION

Rubrics provide a standardized method for evaluating student pharmacists' performance on learning activities.^{7-11,14,22,27} Previous studies have found that students believe using a rubric to grade assignments limits

subjectivity and variability when multiple evaluators are involved in a course.^{5,9,27} As a standardized method of evaluating clinical skills and documentation, rubrics should meet educational standards for reliability.^{11,14,22} Continuous evaluation of rubrics across multiple graders and years within a pharmacy curriculum is especially important.¹⁴

The current study demonstrated good IRR across all clinical documentation assignments when the aforementioned rubric was utilized. Analysis of different subgroups indicated variation in IRR between the student pharmacist's year in the professional program and disease state. Some of the variation in IRR between assignments may have been due to the complexity of the disease and patient case being evaluated. The IRR for evaluations of P3 cases was fair, but lower than that for P2 cases. This may have been related to the difficulty of the topics. However, previous literature has described the practice of accepting lower levels of reliability for open-ended, formative assessments.⁷ Nevertheless, it is still vital to expose student pharmacists to the inherent variability in treatment options within clinical practice through open-ended assignments.

Another reason for the overall good IRR not being stronger may have been the variability in evaluator experience and limited training in using a rubric. First, the evaluators represented diverse levels of clinical and academic experience. Half of the evaluators were pharmacy residents with less than two years of professional experience, while practice faculty members had one to 12 years of clinical and academic experience, which could have impacted the overall IRR. Additionally, training of the evaluators could have impacted the IRR. While all evaluators underwent orientation and training for use of the rubric at the beginning of the semester, training was limited. The program would benefit from evaluators receiving extended training, including longitudinal

Table 2. Clinical Documentation Rubric Scores and Intra-class Correlation Coefficients for Faculty Evaluators

Variable	No.	Score 1, Mean (SD)	Score 2, Mean (SD)	ICC (95% CI) ^a
All students	317	9.1 (0.9)	9.1 (0.9)	0.7 (0.6-0.8) ^a
Year in program				
Second professional	234	9.0 (1.0)	9.0 (1.0)	0.7 (0.7-0.8) ^a
Third professional	83	9.3 (0.7)	9.4 (0.6)	0.5 (0.4-0.7) ^a
Disease state				
Community acquired pneumonia	49	9.2 (0.9)	9.3 (0.8)	0.5 (0.3-0.7) ^a
Contraception	59	9.5 (0.6)	9.5 (0.6)	0.5 (0.3-0.7) ^a
Diabetes	83	8.7 (1.1)	8.7 (1.1)	0.9 (0.8-0.9) ^a
Alcohol withdrawal	92	9.1 (0.8)	8.9 (1.0)	0.5 (0.4-0.7) ^a
Sexually transmitted infections	34	9.5 (0.3)	9.5 (0.4)	0.6 (0.3-0.8) ^a

Abbreviations: ICC=intra-class correlation coefficients, CI=confidence interval

^a All *p*-values <.01

education.²⁷ Nevertheless, these results show that implementation of rubric use by multiple evaluators resulted in good IRR for grading clinical documentation.

The results of this study are consistent with those of previous research evaluating the reliability of rubrics implemented within pharmacy education.^{14,15} Andrus and colleagues reported ICC values for a rubric and noted that with even fair ICC scores it resulted in improved student-submitted SOAP note scores from first to final assignment submission.¹⁵ Sherman and colleagues evaluated the impact of a rubric and standardized patient feedback on longitudinal student pharmacist SOAP note performance.¹⁴ Their study showed enhanced student performance on longitudinal SOAP notes, with the authors pointing to the vital feedback from the rubric.¹⁴ While the current study does not assess individual student performance or perceptions, the other studies within pharmacy education highlight the importance of rubrics for longitudinal student development.

Based on the findings of this study, minor modifications were made to the rubric. With some aspects of the rubric showing only fair reliability, descriptions were added to enhance standardization of grading across evaluators. For example, instead of just utilizing vague terms (e.g., some, partially, mostly), percentages were added (e.g., 75%) for greater precision. Other minor modifications made related to scoring and the inclusion of an automatic “do not pass” for potentially sentinel events. Also, not because of the study findings, but for accreditation purposes, the PPCP was incorporated into the rubric and clearly defined.^{1,3} Finally, moving forward, longitudinal training for evaluators and student training on how the rubric can track skill achievement will be incorporated into the rubric in the future.²²

Limitations exist for this study. One was the variation in the number of rubrics completed for the two years (P2 and P3) of pharmacy students included in the

program. The difference in the number of rubrics completed in each year makes it difficult to compare the reliability of the rubric over time. Another limitation is the lack of overlapping topics across years in the pharmacy program, which again limits the ability to compare the reliability across years in the pharmacy program. Additional limitations include cross-sectional data analysis of a single semester, the use of only two evaluators per submission, the differences in evaluator experience, and the small number of disease states the rubric was applied to, which made it difficult to extrapolate findings across the curriculum. Finally, the study focused on evaluator IRR and did not assess the impact on student learning.

CONCLUSION

Rubrics can be an important component in the delivery and assessment of the pharmacy curriculum. The clinical documentation rubric implemented here demonstrated overall good inter-rater reliability between multiple individuals when used in pharmacotherapy PBL courses. The rubric will continue to undergo evaluation and quality improvement to ensure the provision of necessary formative feedback to student pharmacists.

REFERENCES

1. Joint Commission of Pharmacy Practitioners. Pharmacists' Patient Care Process. May 29, 2014. <https://www.pharmacist.com/sites/default/files/files/PatientCareProcess.pdf>. Accessed July 10, 2020.
2. American Society of Health-System Pharmacists. ASHP guidelines on documenting pharmaceutical care in patient medical records. *Am J Health Syst Pharm*. 2003;60(7):705-707.
3. Accreditation Council for Pharmaceutical Education. Accreditation standards and key elements for the professional program in pharmacy leading to the Doctor of Pharmacy degree (“Standards 2016”). Published February 2, 2015. <https://www.acpe-accredit.org/pdf/Standards2016FINAL.pdf>. Accessed July 10, 2020.
4. Fleming PR, Manderson WG, Matthews MB, Sanderson PH, Stokes JF. Evolution of an examination: M.R.C.P. (U.K.). *Br Med J*. 1974;2(5910):99-107.

5. McManus IC, Thompson M, Mollon J. Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP(UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Med Educ.* 2006;6(1):Article 42.
6. Black PJ. Testing: friend or foe? the theory and practice of assessment and testing. London, England: Routledge; 1997.
7. Jonsson A, Svingby G. The use of scoring rubrics: reliability, validity, and educational consequences. *Educational Research Review.* 2007;2(2):130-144.
8. Tierney R, Simon M. What's still wrong with rubrics: focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research, & Evaluation.* 2004;9(2):Article 2.
9. Phillips JM, Stalter AM, Ruggiero JS, et al. Systems thinking for transitions of care: reliability testing for a standardized rubric. *Clin Nurse Spec.* 2019;33(3):128-135.
10. Andrade HG. Teaching with rubrics: the good, the bad, and the ugly. *College Teaching.* 2005;53(1):27-31.
11. Reddy YM, Andrade H. A review of rubric use in higher education. *Assess Eval High Educ.* 2010;35(4):435-448.
12. Sando KR, Skoy E, Bradley C, Frenzel J, Kirwin J, Urteaga E. Assessment of SOAP note evaluation tools in colleges of pharmacy. *Curr Pharm Teach Learn.* 2017;9(4):576-584.
13. Peeters MK, Sahloff EG. A standardized rubric to evaluate student presentations. *Am J Pharm Educ.* 2010;74(9):Article 171.
14. Sherman JJ, Johnson CD. Assessment of pharmacy students' patient care skills using case scenarios with a SOAP note grading rubric and standardized patient feedback. *Curr Pharm Teach Learn.* 2019;11(5):513-521.
15. Andrus MR, McDonough SLK, Kelley KW, et al. Development and validation of a rubric to evaluate SOAP note writing in APPE. *Am J Pharm Educ.* 2018;82(9):Article 6725.
16. Salinitri FD, Lobkovich AM, Crabtree BL, Wilhelm SM. Reliability and validity of a checklist to evaluate student performance in a problem-based learning group. *Am J Pharm Educ.* 2019;83(8):Article 6963.
17. Barnett SG, Gallimore C, Kopacek KJ, Porter AL. Evaluation of electronic SOAP note grading and feedback. *Curr Pharm Teach Learn.* 2014;6(4):516-526.
18. Horton N, Payne KD, Jernigan M, et al. A standardized patient counseling rubric for a pharmaceutical care and communications course. *Am J Pharm Educ.* 2013;77(7):Article 152.
19. Haack S, Fornoff A, Caligiuri F, et al. Comparison of electronic versus paper rubrics to assess patient counseling experiences in a skills-based lab course. *Curr Pharm Teach Learn.* 2017;9(6):1117-1122.
20. Brown MC, Conway J, Sorensen TD. Development and implementation of a scoring rubric for aseptic technique. *Am J Pharm Educ.* 2006;70(6):Article 133.
21. Blommel ML, Abate MA. A rubric to assess critical literature evaluation skills. *Am J Pharm Educ.* 2007;71(4):Article 63.
22. Peeters MJ. Measuring rater judgments within learning assessments – Part 2: a mixed approach to creating rubrics. *Curr Pharm Teach Learn.* 2015;7(5):662-668.
23. Schwinghammer TL, Koehler JM. *Pharmacotherapy Casebook: A Patient-focused Approach.* 8th ed. New York, NY: McGraw Hill; 2011.
24. UNC Eschelmann School of Pharmacy. SOAP Note Scoring Rubric. <https://faopharmacy.unc.edu/files/2015/05/SOAP-Rubric-October-2012.pdf>. Accessed July 10, 2020.
25. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med.* 2016;15(2):155-163.
26. Hallgren KA. Computing inter-rater reliability for observational data: An overview and tutorial. *Tutor Quant Methods Psychol.* 2012; 8(1):23-34.
27. Andrade H, Du Y. Student perspectives on rubric-referenced assessment. *Practical Assessment Research & Evaluation.* 2005;10(3):1-11.