EDUCATION
RESEARCH SERVICE

ORIGINAL ARTICLE

Development of a student grading rubric and testing for interrater agreement in a doctor of chiropractic competency program

Krista Ward, DC, MPH, Kathy Kinney, DC, Rhina Patania, DC, Linda Savage, DC, Jamie Motley, MS, DC, and Monica Smith, DC, PhD

Objective: Clinical competency is integral to the doctor of chiropractic program and is dictated by the Council of Chiropractic Education accreditation standards. These meta-competencies, achieved through open-ended tasks, can be challenging for interrater agreement among multiple graders. We developed and tested interrater agreement of a newly created analytic rubric for a clinical case-based education program.

Methods: Clinical educators and research staff collaborated on rubric development and testing over four phases. Phase 1 tailored existing institutional rubrics to the new clinical case-based program using a 4-level scale of proficiency. Phase 2 tested the performance of the pilot rubric using 16 senior intern assessments graded by four instructors using pre-established grading keys. Phases 3 and 4 refined and retested rubric versions 1 and 2 on 16 and 14 assessments, respectively.

Results: Exact, adjacent, and pass/fail agreements between six pairs of graders were reported. The pilot rubric achieved 46% average exact, 80% average adjacent, and 63% pass/fail agreements. Rubric version 1 yielded 49% average exact, 86% average adjacent, and 70% pass/fail agreements. Rubric version 2 yielded 60% average exact, 93% average adjacent, and 81% pass/fail agreements.

Conclusion: Our results are similar to those of other rubric interrater reliability studies. Interrater reliability improved with later versions of the rubric likely attributable to rater learning and rubric improvement. Future studies should focus on concurrent validity and comparison of student performance with grade point average and national board scores.

Key Indexing Terms: Education; Chiropractic; Validation Studies as Topic

J Chiropr Educ 2019;33(2):140–144 DOI 10.7899/JCE-18-9

INTRODUCTION

The Council on Chiropractic Education (CCE) 2018 Accreditation Standards dictates that graduates of accredited doctor of chiropractic programs are competent in clinical reasoning and the following eight meta-competencies: Assessment and Diagnosis, Management Plan, Health Promotion and Disease Prevention, Communication and Recording Keeping, Information and Technology Literacy, Chiropractic Adjustment/Manipulation, and Inter-Professional Education.[1] Open-ended tasks, such as free-text written assessments synthesizing a patient's history and exam are needed to elicit students' clinical reasoning and higher order thinking.[2,3] However, assessing these open-ended tasks can be a challenge for the instructor. In secondary and higher education, a rubric typically is used to assess this type of student performance.

Rubrics are tools used to help objectively measure student performance on written assessments. Unlike checklists that detail student requirements, a rubric is "essentially a scaled tool with levels of achievement and clearly defined criteria related to each level and placed in a grid."[4] Grading time also is reduced, since an instructor's repetitive feedback can be incorporated in the rubric criteria.[5] While a plethora of studies exist on rubrics in education, including clinical settings,[2–6] only one conference presentation describes a rubric in a doctor of chiropractic clinical training setting.[7]

The Clinical Education Department at our institution has a 10-year history of applying a rubric to grade intern performance on a case-based proficiency exam taken midway through the clinical experience and a 2-year history of using a rubric for our case-based radiology program, called "radiology case of the week" (RCW), wherein students practice and demonstrate their clinical skills in writing a report based on new imaging films they receive each week. The RCW grading rubric currently assesses a student's skillfulness along eight distinct

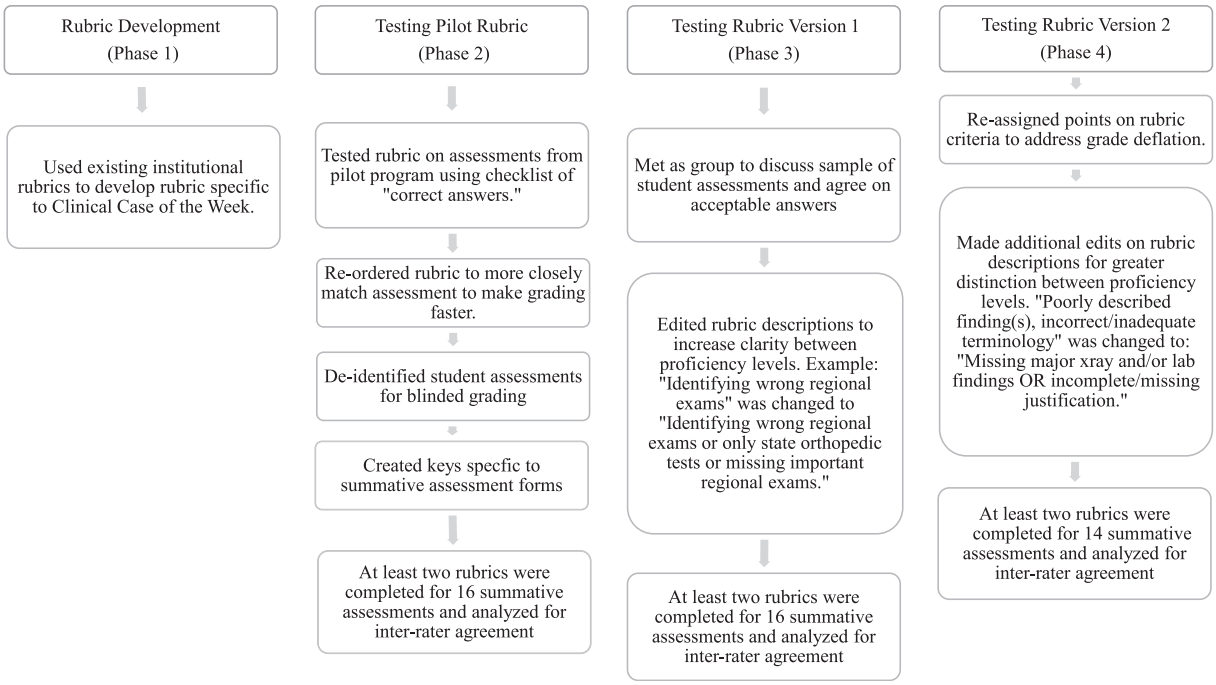

Figure 1 - Phases of rubric development and testing.

dimensions of radiology report writing, with the student's performance on each dimension being assigned into one of four discrete levels of proficiency (inadequate, novice, competent and proficient). For example, one tested dimension is the ability of students to provide appropriate recommendations. A student performing at the proficient level will include in their report the "Most appropriate test, lab, additional imaging requested with justification." A student performing at the competency level meets the requirement for proficiency, except they do not include the justification and a student performing at the inadequate level will not recommend necessary diagnostic studies or referrals or they will request an unnecessary referral.

In 2017, we developed and introduced a new "clinical case of the week" (CCW) program for students to practice and demonstrate their diagnostic and case management skills (CCE meta-competencies 1 and 2). In CCW, students are given information from an example patient record in a 4-step staged presentation: First, the patient demographics and pain drawing are presented; second, presentation of patient subjective findings; third, physical exam findings; and finally, radiology and laboratory reports. After each step, students are expected to answer questions in free-text format to demonstrate their competency in clinical case-based assessment, diagnosis and management planning. Our institution's RCW 4 proficiency level rubric was used as a template for a new CCW rubric to grade these free-text student responses. We report our experience with developing and testing the CCW analytic rubric.

METHODS

Phase 1: Rubric Development

Three clinical educators with collectively 60 plus years of clinical and teaching experience at the institution worked with research department staff to develop and test a grading rubric for CCW. Figure 1 outlines the timeline of the four academic terms (phases 1–4) of this project.

In phase 1, we used the already established rubrics from the proficiency exam and the RCW as starter templates to begin our development of the CCW rubric. During this phase, we continued refining the CCW rubric to explicate important dimensions within clinical reasoning skillsets and to better align each question on the CCW assessment form with the identified dimensions. The 17 final rubric dimensions representing the expectations of the CCW assessment are listed in Figure 2.

Phase 2: Rubric Pilot-Test and Baseline Interrater Agreement

In phase 2 of the CCW program roll-out, each clinician attempted to grade an assessment using the rubric. Through the process of grading, changes to the rubric and process were identified including de-identifying the assessments to reduce bias, reordering the rubric dimensions to more closely match the order of assessment questions, and create detailed keys specific to each assessment (Fig. 1). We then tested the pilot rubric for interrater agreement using 16 assessments completed by senior interns in the new CCW Program.

- Step 1: (a) Chief Complaint; (b) Possible Diagnoses; (c) Questions
- Step 2: (d) Modifiers; (e) Most likely condition considered with subjective findings; (f) DDX1 considered with subjective findings; (g) DDX2 considered with subjective findings; (h) Regional Exam Selection with justification
- Step 3: (i) Most likely condition ruled in with objective findings; (j) DDX1 ruled out with objective findings; (k) DDX2 ruled out with objective findings; (l) Imaging, Laboratory and/or Specialist Referral with justification
- Step 4: (m) Indications from imaging and laboratory findings; (n) Problems List Details; (o) Case Management Proposal including contraindications and referrals; (p) Recommended Outcome Assessment Tools and Screens; (q) Prognosis supported by case details

Figure 2 - Four-step staged presentation of clinical case of the week, with 17 (a–q) dimensions of clinical reasoning skills. DDX = differential diagnosis

The grading task was divided among four team members (KK, RP, LS, KW). After initial grading was complete, the 16 assessments were exchanged among the graders and regraded independently. One assessment was independently graded by three examiners, creating 18 rubric pairs that then were compared for exact and adjacent agreement (agreement within 1 proficiency level). Exact "pass/fail" agreement between graders also was assessed using only two proficiency levels (competent and proficient were combined into "pass" and inadequate and novice combined into "fail"). The data were entered into Microsoft Excel (Microsoft Corp, Redmond, WA) for calculations of average agreement and standard errors.

Phases 3–4: Rubric Refinement and Improved Interrater Agreement

Following the pilot testing, additional revisions were made to the rubric to increase clarity between proficiency levels. The revised rubric (V1) was tested on 16 assessments from a new cohort of interns. Six assessments were independently graded by multiple team members creating 30 pairs for reliability testing. In the last phase, the rubric was refined further by reassigning points so that students performing at a competent level would meet a 75-point passing threshold. In the previous versions, if a student scored competent on each of the 17 dimensions, they would only have earned 70 points. V2 was tested on 14 assessments, with 18 graded pairs.

RESULTS

Table 1 shows our team averages and standard errors for the calculated percentages of exact, adjacent, and pass/fail agreements between the six pairs of graders (KW and KK, KW and RP, KW and LS, KK and RP, KK and LS, LS and RP).

With each subsequent version of the rubric, we achieved improvements in all calculated averages for interrater agreement (Table 1), eventually attaining >90% average adjacent agreement on a 4-level scale of proficiency scoring.

DISCUSSION

A 2007 review of educational rubrics by Jonsson and Svingby found that the percentage of exact agreement varied among studies of interrater reliability with the majority of estimates falling <70%, which as cited by Stemler in the review by Johnsson and Svingby "is needed if exact agreement is to be considered reliable."[3] Jonsson and Svingby also noted that rater agreement depended on the number of levels in the rubric.[3] A study of a rubric

Table 1 - Average Percentage of Exact, Adjacent, and Pass/Fail Agreement with Standard Errors for each Rubric Version

Rubric Version	Number of Graded Assessments Pairs	Average Agreement (SE)		
		Exact	Adjacent	Pass/Fail
Pilot	18	46% (.038)	80% (.031)	63% (.041)
V1	30	49% (.025)	86% (.021)	70% (.026)
V2	18	60% (.024)	93% (.011)	81% (.021)

assessing medicine core clerkship write-ups with 14 items and a 4-point scale reported a median of 54% exact agreement and 94% adjacent agreement.[8] These results are very similar to our own of 60% and 93% exact and adjacent agreement, respectively. Similar to other studies, our percentage of adjacent agreement was much higher than our percentage of exact agreement and was >90%, which Jonsson and Svingby stated is "a good level of consistency."[3]

One limitation of our study is the possibility that chance agreement overestimated our results. For instance, our graders sometimes marked in between proficiency levels on the rubric or marked more than 1 level if they were unsure. McHugh notes "if there is likely to be much guessing among the raters, it may make sense to use the kappa statistic, but if raters are well trained and little guessing is likely to exist, the researcher may safely rely on percent agreement to determine interrater reliability."[9] Our sample sizes were too small to use inferential testing of interrater reliability using the kappa statistic. McHugh notes that " . . . as a general heuristic, sample sizes should not consist of less than 30 comparisons. . .."[9] While we had 30 comparisons for V1, these data were from only 16 student assessments with multiple examiner pairs.

Another limitation of our study is decreased generalizability given that the study was done at a single institution, and all faculty raters were well familiarized with the cases. As found by others,[3] we observed the importance of standardizing the training of graders to ensure more consistent application of the grading rubric. One explanation for the higher agreement found with the final rubric version was that we spent more time discussing the case and developing an in-depth answer key before using it to grade. Reliability may not be as high among graders who are either unfamiliar with the case or who do not participate in developing the answer key.

Finally, by providing only percent agreements for the rubrics as a whole, we do not know how often graders agreed or disagreed on each individual dimension. Some dimensions are more subjective than others (for example, case management plan compared to modifiers) and agreement on relatively straightforward dimensions may have artificially elevated the use of the rubric for grading subjective responses.

CONCLUSION

For consistency with the radiology rubric, we used the same four levels to categorize an intern's clinical skill proficiency in CCW. To "pass" any given weekly case a student's clinical skill must be considered "competent" or "proficient." Students are required to attain three Competent/Proficient summative assessments for graduation and a reliable rubric is needed to make high stake decisions.[3] Our results are consistent with those of other rubric interrater reliability studies and, given the multiple dimensions and four scales of our CCW rubric, 81% pass/fail agreement is considered reliable. Future studies should focus on concurrent validity and compare student performance on the CCW rubric with GPA and national board of chiropractic examiner NBCE scores.

ACKNOWLEDGMENTS

The authors thank Barbara Delli Gatti, MLS, MSEd for acquiring relevant research and preparing the citations for this manuscript, contributing to integrating the peer-reviewed research into the introduction, and editing the final document.

FUNDING AND CONFLICTS OF INTEREST

This project received no funding and the authors have no conflicts of interest to declare.

About the Authors

Krista Ward is a research specialist and adjunct faculty member of the Research Department at Life Chiropractic College West (25001 Industrial Blvd, Hayward, CA 94545; kward@lifewest.edu). Kathy Kinney is a professor in the health center at Life Chiropractic College West (25001 Industrial Blvd, Hayward, CA 94545; kkinney@lifewest.edu). Rhina Patania is a professor in the health center at Life Chiropractic College West (25001 Industrial Blvd, Hayward, CA 94545; rpatania@lifewest.edu). Linda Savage is a professor in the health center at Life Chiropractic College West (25001 Industrial Blvd, Hayward, CA 94545; lsavage@lifewest.edu). Jamie Motley is an associate professor in the Radiology Department at Life Chiropractic College West (25001 Industrial Blvd, Hayward, CA 94545; jmotley@lifewest.edu). Monica Smith is the Director of Research in the Research Department at Life Chiropractic College West (25001 Industrial Blvd, Hayward, CA 94545; MSmith@lifewest.edu). Address correspondence to Krista Ward, Life Chiropractic College West, 25001 Industrial Blvd, Hayward, CA 94545; kward@lifewest.edu.

This article was received April 23, 2018, revised September 4, 2018, and accepted October 30, 2018.

Author Contributions

Concept development: MS, LS, KW, KK, RP, JM. Design: MS, LS, KW, KK, RP, JM. Supervision: MS, KW. Data collection/processing: LS, KW, RP, KK. Analysis/interpretation: KW. Literature search: MS. Writing: MS, JM, KW. Critical review: MS, LS, KW, KK, RP, JM.

REFERENCES

1. Council on Chiropractic Education (CCE). *CCE Accreditation Standards: Principles, Processes & Requirements for Accreditation*. Scottsdale, AZ: Council on Chiropractic Education; 2018.

2. Durning SJ, Artino A, Boulet J, et al. The feasibility, reliability, and validity of a post-encounter form for evaluating clinical reasoning. *Med Teach*. 2012;34(1): 30–37.
3. Jonsson A, Svingby G. The use of scoring rubrics: reliability, validity and educational consequences. *Educ Res Rev*. 2007;2(2):130–144.
4. O'Donnell JA, Oakley M, Haney S, O'Neill PN, Taylor D. Rubrics 101: a primer for rubric development in dental education. *J Dent Educ*. 2011;75(9):1163–1175.
5. Isaacson JJ, Stacy AS. Rubrics for clinical evaluation: objectifying the subjective experience. *Nurse Educ Pract*. 2009;9(2):134–140.
6. Orrock P, Grace S, Vaughan B, Coutts R. Developing a viva exam to assess clinical reasoning in pre-registration osteopathy students. *BMC Med Educ*. 2014;14:193.
7. Ciolfi M. Clinical competencies assessment rubric system (C-CARS). Proceedings of the 2009 Association of Chiropractic Colleges Educational Conference XVI and Research Agenda Conference XIV; 2009 Mar 13-14. *J Chiropr Educ*. 2009;23(1):62.
8. Kogan JR, Shea JA. Psychometric characteristics of a write-up assessment form in a medicine core clerkship. *Teach Learn Med*. 2005;17(2):101–106
9. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)*. 2012;22(3):276–282.