RESEARCH

Determination of Interrater Reliability of a Universal Evaluator Rubric to Assess Student Pharmacist Communication Skills

Susanne G. Barnett, PharmD,a Sheila M Allen, PharmD,b Karen MS Bastianelli, PharmD,c Jennifer S. Chen, PharmD,c Colleen A. Clark Dula, PharmD,d Marlowe Djuric Kachlic, PharmD,b Kristen L. Goliak, PharmD,b Laura E. Knockel, PharmD,e David E. Matthews, PharmD,d Lucio R. Volino, PharmD,f Michael R. Lasarev, MS,a Jeffrey C. Reist, PharmD,e

a University of Wisconsin-Madison, School of Pharmacy, Madison, Wisconsin
b University of Illinois at Chicago, College of Pharmacy, Chicago, Illinois
c University of Minnesota, College of Pharmacy, Minneapolis, Minnesota
d The Ohio State University, College of Pharmacy, Columbus, Ohio
e University of Iowa, College of Pharmacy, Iowa City, Iowa
f Rutgers University, Ernest Mario School of Pharmacy, Piscataway, New Jersey

Submitted January 9, 2021; accepted June 11, 2021; ePublished June 2021

Objective. To evaluate the interrater reliability of a universal evaluator rubric used to assess student pharmacist communication skills during patient education sessions.

Methods. Six US schools and colleges of pharmacy each submitted 10 student videos of a simulated community pharmacy patient education session and recruited two raters in each of the five rater groups (faculty, standardized patients, postgraduate year one residents, student pharmacists, and pharmacy preceptors). Raters used a rubric containing 20 items and a global assessment to evaluate student communication of 12 videos. Agreement was computed for individual items and overall rubric score within each rater group, and for each item across all rater groups. Average overall rubric agreement scores were compared between rater groups. Agreement coefficient scores were categorized as no to minimal, weak, moderate, strong, or almost perfect agreement.

Results. Fifty-five raters representing five rater groups and six pharmacy schools evaluated student communication. Item agreement analysis for all raters revealed five items with no to minimal or weak agreement, 10 items with moderate agreement, one item with strong agreement, and five items with almost perfect agreement. Overall average agreement across all rater groups was 0.73 (95% CI, 0.66-0.81). The preceptor rater group exhibited the lowest agreement score of 0.68 (95% CI, 0.58-0.78), which significantly deviated from the overall average.

Conclusion. While strong or almost perfect agreement scores were not observed for all rubric items, overall average interrater reliability results support the use of this rubric in a variety of raters to assess student pharmacist communication skills during patient education sessions.

Keywords: communication, medication education, patient education, assessment tool, interrater reliability

INTRODUCTION

Communication is a key element necessary to solve problems, educate, advocate, and collaborate within health care. The Accreditation Council for Pharmacy Education (ACPE) Standards 2016 requires that Doctor of Pharmacy (PharmD) graduates be able to verbally and nonverbally communicate with individuals, groups, and organizations.1 Communication is also emphasized as a central part of the Pharmacists’ Patient Care Process (PPCP).2 Pharmacists collaborate with patients, caregivers, and other health care providers to communicate effectively and provide care in each step of the PPCP. Effective pharmacist communication can increase medication adherence and reduce medication errors.3,4

As such, schools and colleges of pharmacy provide learners with ample instruction and opportunities to practice communication skills within their curricula. To ensure minimum competence and progression towards advanced pharmacy practice experiences (APPEs),
consistent performance-based assessments and reliable evaluation tools, such as communication rubrics, are needed. A review of the literature found that various teaching strategies are used to accomplish this imperative task.5,6 There are also numerous rubrics that have been proposed to assess communication skills within various environments.7-11 To our knowledge, no study has used coordinated evaluator training with a standardized codebook in an effort to increase the reliability of a communication evaluation tool. Previously published studies have described methods for testing and achieving interrater reliability; however, these were limited to single evaluator types.11,12 There is no rubric for assessment of communication skills among student pharmacists that has been validated for use among multiple evaluator types (ie, senior student colleagues, residents, faculty, preceptors, and/or standardized patients [SPs]).

In recognition of the challenges and variability associated with evaluating student pharmacists’ communication skills, members of the Big Ten Academic Alliance (BTAA) Performance-Based Assessment Collaborative (PBAC) united to build expert consensus and create a universal evaluator rubric (for use among multiple evaluator types) with standardized training to assess student pharmacist communication skills.13 The goal of this study was to evaluate the interrater reliability of the communication rubric developed by the BTAA-PBAC among multiple rater groups during a patient education encounter.

METHOD

Creation of a 20-item universal evaluator communication rubric and corresponding codebook was previously described in a multi-step process of expert consensus building involving faculty and patient content experts.13 The rubric also asked raters to score a yes/no global assessment of whether they would return to the student pharmacist. The codebook was created and refined using a two-phase process in order to increase rater consistency when scoring student communication. Using this communication rubric, skills laboratory faculty and instructors from six BTAA-PBAC member schools and colleges (University of Wisconsin, University of Iowa, Rutgers University, University of Minnesota, The Ohio State University, and University of Illinois at Chicago) collaborated to determine interrater reliability in five rater groups (pharmacy school/college faculty, standardized patients, postgraduate year one [PGY1] residents, student pharmacists, and pharmacy preceptors [Table 1]).

Each of the six participating schools and colleges acted as a site for rater recruitment and training and submitted 10 de-identified student videos of simulated community pharmacy patient education sessions. Videos were chosen at random from consenting students who were enrolled in a skills-based laboratory course. Education sessions submitted by individual schools varied in consultation type (one new, one new and one refill, and two refill prescriptions), duration (3-15 minutes), and year of student (first, second, and third year PharmD students). However, all 10 videos submitted by a given school were taken from the same simulated activity. Each skills laboratory course required students to participate in the patient education session, and video recordings were completed as a part of normal course operations. A total of 60 student videos were submitted to a central repository housed at the University of Wisconsin-Madison.

Two raters from each of the five rater groups were recruited by the six participating schools and colleges. In the event that a school did not use a rater group to evaluate student performance (eg, a school did not use SPs in their program), that school did not recruit for the rater group in question. Prior to study participation, raters were required to attend a standardized, 90-minute, in-person training and successfully pass a certification process in an effort to maximize intra-rater reliability during the study. Training materials included a prerecorded video that discussed each of the 20 communication rubric items and scoring according to the codebook. The training video also showed two counseling sessions, each of a student providing education to an SP. After the first student clip, the video discussed how each rubric item should be scored while referencing the codebook. After the second student clip, the video was paused and raters were asked to use the communication rubric and codebook to evaluate the student on their own. When the training video was restarted, the scoring of each rubric item was discussed. The training concluded with a certification process in which a rater was given two opportunities to score a student education session and obtain a minimum score of 80% agreement with a key. Keys were developed through consensus of pharmacy practice faculty at the six schools and colleges participating in this study. Faculty scored the patient education sessions independently using the rubric and met to discuss and come to consensus on any items not rated consistently across all faculty. Raters were not invited to participate in the study if certification had not been obtained.

If a rater did not pass certification, their school recruited a rater to fill the initial spot only once. For example, if one of the two SP raters at a given school failed certification, a third SP would be recruited and trained. If this third SP rater failed certification, additional SPs from that school were not recruited and only one SP from the school would participate in the study. All raters passing certification were given 28 days from the date of training to...
evaluate student communication skills on a total of 12 videos, including the 10 videos originating from their school. Raters from different universities but similar rater groups (eg, faculty) were randomly paired and assigned to evaluate two additional videos from another school to ensure rater reliability across different consultation types. An overlap of 20% was chosen in order to assess and control for possible bias between raters.14

Raters returned completed paper communication rubrics to study coordinators, who entered all data into Qualtrics (Qualtrics International). Gwet’s agreement coefficient was used to quantify chance-corrected agreement among several raters scoring multi-level (“no,” “inconsistent,” and/or “yes”) items.15,16 Gwet’s agreement coefficient was chosen because of its ability to avoid paradoxes that plague kappa (eg, a counter-intuitive high percentage of agreement and low kappa value when prevalence is rare). Agreement was computed for each item and each of the five rater groups; an overall measure of agreement for each item was computed using all data without concern for rater group. While no uniform guideline for interpretation of Gwet’s agreement coefficient exists, interpretation can be generalized to that of kappa, a less robust agreement coefficient that may produce paradoxical conclusions if responses in a category are rare.17-19

Based on recommendations for interpretation of Gwet’s agreement coefficient scores for health care-related studies, agreement scores <0.39 were interpreted as no to minimal agreement, 0.4-0.59 as weak agreement, 0.6-0.79 as moderate agreement, 0.8-0.9 as strong agreement, and >0.90 as almost perfect agreement.20 Standard errors for each agreement score were estimated, and approximate 95% confidence intervals for agreement were formed as the estimate ±2 standard errors. An informal analysis was done to estimate differences among rater groups with respect to average overall agreement from items 1-20 and excluded the global assessment item. Agreement scores for each item and rater group were analyzed using a generalized estimating equation in which individual items evaluated by a rater group were treated as the “panel.”21 Analyses were done using R, version 4.0.2 (R Core Team, Vienna, Austria). This study was approved by the institutional review board at each of the six participating schools and colleges of pharmacy.

RESULTS

Participant recruitment, exclusion, and inclusion by rater group are summarized in Table 2. Out of 58 raters initially recruited and trained, six did not pass certification and were excluded from analysis. Half of these were student pharmacist raters. Two raters did not evaluate video recordings within the required 28-day window and were also excluded from analysis. Of the eight raters who did not meet inclusion criteria, five were re-recruited, certified, and included in the analysis. Ultimately, 55 raters over five rater groups were included in the analysis.

Rubric item agreement scores for each of the five rater groups and overall are summarized in Table 3 and Figure 1. Item agreement analysis for all raters revealed five items with no to minimal or weak agreement, 10 items with moderate agreement, one item with strong agreement, and five items with almost perfect agreement. The rubric items with the highest (almost perfect) agreement scores across all rater groups included “used teach-back” (item 18), “conveyed non-judgmental attitude” (item 5), “conveyed respect for patient” (item 4), “student introduced him/herself by name and title” (item 1), and “allowed patient to speak without interruption” (item 10). Rubric items with the lowest (no to minimal or weak) agreement scores across all rater groups included “avoided filler words” (item 9), “achieved mutual agreement with the plan” (item 19), “provided opportunity for and responded to questions” (item 17), “established purpose of

Table 1. Rater Groups Included in a Study to Determine the Interrater Reliability of a Universal Evaluator Rubric to Assess Student Pharmacist Communication Skills

| Rater Group          | Definition                                                                 |
|----------------------|-----------------------------------------------------------------------------|
| Faculty              | A professor of any ranking, or a fellow or post-graduate year-2 resident focusing on teaching in the laboratories. |
| Standardized patient | An individual acting in the role of a patient (often with a script) in an attempt to provide a standardized experience to multiple students. |
| Postgraduate year 1 resident | An individual with a Doctor of Pharmacy degree who is enrolled in a post-graduate year-1 resident program. |
| Student pharmacist   | Any individual currently enrolled in a Doctor of Pharmacy program. |
| Preceptor            | A pharmacist who serves as an instructor for Doctor of Pharmacy students at an experiential site. |
Table 2. Recruitment, Exclusion, and Inclusion of Participants Included in a Study to Determine the Interrater Reliability of a Universal Evaluator Rubric to Assess Student Pharmacist Communication Skills

| Participant                      | Initially recruited | Excluded from Analysis | Included in Analysis |
|----------------------------------|---------------------|------------------------|----------------------|
| Faculty<sup>a</sup>              | 11                  | –                      | –                    |
| Standardized patient<sup>b</sup>| 10                  | 2 1                    | 2                    |
| Postgraduate year 1 resident     | 12                  | –                      | 1                    |
| Student pharmacist<sup>c</sup>   | 14                  | 3 –                    | 1                    |
| Preceptor<sup>a</sup>            | 11                  | 1 –                    | 1                    |
| All raters                       | 58                  | 6 2                    | 5                    |

<sup>a</sup> One pharmacy school was able to recruit one versus two raters in the faculty and preceptor rater groups.

<sup>b</sup> Standardized patients were utilized and recruited by five of the six participating pharmacy schools.

<sup>c</sup> One pharmacy school initially recruited two additional student raters.

Table 3. Item Rater Agreement Scores on a Communication Rubric to Assess Student Pharmacist Communication Skills

| Communication Rubric Item                                                                 | Agreement Score<sup>a</sup> | Standardized patient | Faculty | Student | Preceptor | Resident | All raters |
|------------------------------------------------------------------------------------------|------------------------------|-----------------------|---------|---------|-----------|----------|------------|
| 1. Student introduced him/herself by name and title                                       | 0.91                         | 0.92                  | 0.95    | 1.00    | 0.88      | 0.93     |
| 2. Elicited concern(s)<sup> (from patient perspective)</sup>                               | 0.57                         | 0.52                  | 0.53    | 0.32    | 0.88      | 0.62     |
| 3. Established purpose of the encounter<sup> (student perspective)</sup>                   | 0.44                         | 0.62                  | 0.55    | 0.54    | 0.60      | 0.54     |
| 4. Conveyed respect for patient                                                          | 0.95                         | 0.97                  | 0.98    | 0.74    | 1.00      | 0.94     |
| 5. Conveyed non-judgmental attitude                                                      | 1.00                         | 0.87                  | 0.97    | 0.99    | 0.96      | 0.95     |
| 6. Listened to and engaged with patient                                                  | 0.82                         | 0.83                  | 0.95    | 0.83    | 0.84      | 0.83     |
| 7. Empathetically responded to patient’s concerns and feelings                           | 0.71                         | 0.61                  | 0.87    | 0.61    | 0.75      | 0.75     |
| 8. Spoke at appropriate pace, volume, and tone                                           | 0.69                         | 0.78                  | 0.63    | 0.59    | 0.69      | 0.68     |
| 9. Avoided filler words (um, uh, ah)                                                     | 0.47                         | 0.60                  | 0.41    | 0.25    | 0.30      | 0.34     |
| 10. Allowed patient to speak without interruption                                         | 0.97                         | 0.98                  | 0.98    | 0.89    | 0.89      | 0.92     |
| 11. Spoke clearly and confidently                                                       | 0.69                         | 0.55                  | 0.65    | 0.66    | 0.65      | 0.64     |
| 12. Used patient friendly language                                                       | 0.66                         | 0.55                  | 0.66    | 0.47    | 0.66      | 0.55     |
| 13. Made appropriate eye contact                                                         | 0.82                         | 0.59                  | 0.82    | 0.82    | 0.67      | 0.73     |
| 14. Displayed appropriate body language                                                  | 0.88                         | 0.83                  | 0.83    | 0.75    | 0.76      | 0.79     |
| 15. Completed one topic prior to moving to the next                                       | 0.88                         | 0.70                  | 0.84    | 0.77    | 0.83      | 0.75     |
| 16. Utilized time efficiently                                                           | 0.65                         | 0.74                  | 0.72    | 0.79    | 0.85      | 0.75     |
| 17. Provided opportunity for and responded to questions                                  | 0.81                         | 0.36                  | 0.44    | 0.72    | 0.64      | 0.47     |
| 18. Used teach-back                                                                      | 0.93                         | 0.98                  | 0.96    | 0.95    | 1.00      | 0.97     |
| 19. Achieved mutual agreement with the plan                                               | 0.65                         | 0.65                  | 0.32    | 0.15    | 0.45      | 0.39     |
| 20. Provides closure to encounter                                                        | 0.84                         | 0.76                  | 0.89    | 0.77    | 0.76      | 0.79     |
| Global (G) Assessment: Would you return to this pharmacist for future guidance on medications? | 0.68                         | 0.85                  | 0.92    | 0.78    | 0.82      | 0.77     |

<sup>a</sup> Agreement score was calculated using Gwet’s unweighted agreement coefficient. Scores <0.39 were interpreted as no to minimal agreement, 0.4-0.59 as weak agreement, 0.6-0.79 as moderate agreement, 0.8-0.9 as strong agreement, and >0.90 as almost perfect agreement.
the encounter” (item 3), and “used patient friendly language” (item 12).

Rater group agreement scores are summarized and compared in Table 4. The overall average agreement across all rater groups was 0.73 (95% CI, 0.66-0.81), which did not differ significantly from average agreement scores of the faculty, SP, postgraduate year 1 resident, and student pharmacist rater groups. The SP rater group exhibited the highest agreement score of 0.77 (95% CI, 0.70-0.84) while the preceptor rater group exhibited the lowest agreement score of 0.68 (95% CI, 0.58-0.78), significantly deviating from the overall average ($p = .03$).

**DISCUSSION**

This study evaluated the interrater reliability of a communication rubric in multiple rater groups during patient education encounters. Previous studies have demonstrated the challenge of developing a communication rubric with a high level of interrater reliability.$^{10,11}$ The Patient-Centered Communication Tools (PaCT) developed by Grice and colleagues demonstrated a significant correlation in interrater scores on two out of five assessment tools but no significant correlation on overall scores. The authors concluded that the interrater scores were not strongly correlated.$^{10}$ Similarly, McKeirnan and colleagues concluded that the majority of
elements on their rubric demonstrated a low level of inter-rater agreement. The current study expands upon the findings from these studies by incorporating recommendations to improve interrater reliability. Previous authors emphasized the importance of training evaluators, with Grice and colleagues specifically recommending the use of multiple videos. Rater subjects in the current study were required to watch and evaluate at least three standardized videos. A certification process in which raters achieved a minimum agreement score (80% or better agreement with the key) on a training video was required for study participation. Additionally, Grice and colleagues also recommended creation of a descriptive rubric. During the expert validation process, a descriptive codebook was developed to accompany the rubric in an effort to decrease subjectivity. Additionally, a key difference in the current study relative to previously published work is the inclusion of multiple rater types rather than only faculty raters. The inclusion of five rater groups allowed the investigation of any differences that may result from variations in rater background.

While the average agreement score for the entire rubric ranged from 0.68-0.77 for each of the five rater groups, SPs achieved the highest level of agreement on the overall rubric score, with at least moderate agreement on 17 items and strong agreement on 11 items. When assessing communication skills, the standardized patient evaluation may be as valuable, if not more valuable, than that of a faculty member or preceptor as it provides an authentic patient perspective. The results of the current study suggest that this rubric can be used by SPs when appropriately trained. In contrast, preceptors had the lowest level of agreement, which may be explained by several factors, including lack of experience with the rubric compared to the other rater groups, focus on the clinical aspects of the encounter rather than communication skills, or a global/practical approach to evaluating communication vs a detailed point-based assessment. For this reason, the rubric may be more appropriate for use in the didactic setting than in the experiential setting. Additionally, prior to use in the experiential settings, schools may consider modification of the codebook for items with minimal to weak agreement, additional training, and/or formative use of the rubric with preceptors in experiential settings.

Agreement across all rater groups for specific rubric items varied widely which may be explained by the subjective nature of some assessment elements. Items with strong to almost perfect agreement included items such as “student introduced him/herself by name and title” (item 1) and “used teach-back” (item 18). Both of these assess communication elements, which tend to be less subjective and easier to ascertain achievement. Other items with strong to almost perfect agreement, including “conveyed respect for patient” (item 4), “conveyed non-judgmental attitude” (item 5), “listened to and engaged with patient” (item 6), and “allowed patient to speak without interruption” (item 10), and “allowed patient to speak without interruption” (item 10), tend to refer to communication elements that are more clearly defined by societal norms and thus less subjective.

The item with the lowest agreement was “provided filler words” (item 9). While filler words may be distracting to some individuals, the ubiquitous use of filler words in communication may make such words less noticeable to those who regularly use filler words in their speech pattern. This could contribute to this item’s subjectivity, resulting in less agreement among raters. Refinement of this item to focus on the use of incongruent filler words, for example, a student using the word “perfect” in their response after a patient states they have diabetes, may make this item more effective at identifying inappropriate filler words and less subjective.

Another item with weak agreement was “provided opportunity for and responded to questions” (item 17). Of note, among rater groups, there was high agreement among SPs on this item while faculty raters had very low agreement. This may represent the tendency of faculty to assess the appropriateness of the student’s clinical response which was not intended to be evaluated in the communication rubric. In contrast, SPs may have simply assessed whether an opportunity for questions was included. To improve agreement on this item, the codebook could explicitly state that the appropriateness of the student’s clinical response should not be assessed using this tool. This should be assessed through other means along with other clinical content. In addition, the dual component (opportunity provided and open-ended nature) may have contributed to lack of agreement. Consideration of both components to assess item 17 is clearly outlined in the codebook; however, it should also be a point of emphasis during rater training.

The option of a rater selecting “non-applicable” for any given item could have been a reason for the observed minimal agreement seen with “achieved mutual agreement with the plan” (item 19). Some of the video scenarios may not have required “agreement with the plan,” which may have been easier for certain rater groups to identify than others. For this item, SPs and faculty had higher agreement than other groups, with preceptors having the lowest agreement. Some might argue that this item does not capture a communication skill but rather a clinical skill and should be assessed in a different manner.

A major strength of the current study is the universal nature of the rubric in regard to its use in multiple evaluator groups. The rubric’s interrater reliability was evaluated using a variety of raters, learners, and patient encounters.
across multiple pharmacy schools. This increases the utility of the rubric to assess student pharmacist communication in a variety of patient encounters. Regardless of resources available to a school and the need to transition the assessment of student communication from one rater type to another, continued use of the rubric is justified. Despite these strengths, potential limitations of the current study include that other measures of validity besides inter-rater reliability were not assessed, and that we did not capture prior familiarity of the rubric in raters recruited for the study. Future research using this rubric may wish to explore test-retest reliability, and the performance of the rubric in the experiential setting. Despite inclusion of multiple institutions to increase generalizability in the current study, these institutions were limited to Big Ten schools.

**CONCLUSION**

While strong or almost perfect agreement scores were not observed for all rubric items evaluated, overall average interrater reliability results support the use of this rubric in a variety of raters to assess student communication during patient education sessions. Additional research should be aimed at further refining rubric items that performed poorly and/or considering rubric items that may need to be the focus of additional training during the rater certification process.

**REFERENCES**

1. Accreditation Council for Pharmacy Education. Accreditation standards and key elements for the professional program in pharmacy leading to the Doctor of Pharmacy degree. February 2, 2015. Chicago, Illinois. Accessed January 9, 2022. https://www.acpe-accredit.org/pdf/Standards2016FINAL.pdf
2. Joint Commission of Pharmacy Practitioners. The Pharmacists’ Patient Care Process. Accessed January 9, 2022. https://jcpp.net/wp-content/uploads/2016/03/PatientCareProcess-with-supporting-organizations.pdf
3. De Young M. Research on the effects of pharmacist-patient communication in institutions and ambulatory care sites, 1969-1994. *Am J Health Syst Pharm.* 1996; 53:1277-1291.
4. Davis NM, Cohen MR. Counseling reduces dispensing accidents. *Am Pharm.* 1992;NS32:22.
5. Kimberlin CL. Communicating with patients: skills assessment in US colleges of pharmacy. *Am J Pharm Educ.* 2006;70(3):67.
6. Wallman A, Vaudan C, Sporrong S. Communications training in pharmacy education, 1995-2010. *Am J Pharm Educ.* 2013;77(2): Article 36.
7. Grice G, Gattas N, Sailors J, et al. Health literacy: use of the Four Habits Model to improve student pharmacists’ communication. *Patient Educ Couns.* 2013;90:23-28.
8. Schwartzman E, Hsu D, Law A, Chung E. Assessment of patient communication skills during OSCE: Examining effectiveness of a training program in minimizing inter-grader variability. *Patient Educ Couns.* 2011; 83:472-477.
9. Horton N, Payne K, 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.
10. Grice GR, Gattas NM, Prosser T, et al. Design and validation of Patient-Centered Communication Tools (PaCT) to measure students’ communication skills. *Am J Pharm Educ.* 2017;81(8):Article 5927.
11. McKieran KC, Willson MN, Bray BS. Striving for interrater reliability with a patient counseling assessment rubric: lessons learned. *Curr Pharm Teach Learn.* 2020;12:603-613.
12. Villa KR, Sprunger TL, Walton AM, et al. Inter-rater reliability of clinic documentation rubric within pharmacotherapy problem-based learning courses. *Am J Pharm Educ.* 2020;84(7):Article 7648.
13. Barnett SG, Porter AL, Allen SM, et al. Expert consensus building and creation of a universal evaluator rubric to assess student communication skills. *Am J Pharm Educ.* 2020;84(9):Article 848016.
14. Hallgren KA. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutor Quant Methods Psychol.* 2012; 8(1):23-34.
15. Gwet KL. Computing inter-rater reliability and its variance in the presence of high agreement. *Br J Math Stat Psychol.* 2012; 8(1):23-34.
16. Gwet KL. *Handbook of Inter-Rater Reliability.* 4th ed. Advanced Analytics, LLC; 2014.
17. Rosner B. *Fundamentals of Biostatistics.* 7th ed. Brooks/Cole; 2011.
18. Fleiss JL. *Statistical Methods for Rates and Proportions.* 2nd edition. New York: John Wiley & Sons; 1982.
19. Altman DG. *Practical Statistics for Medical Research.* Chapman & Hall/CRC; 1991.
20. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb).* 2012;22(3):276-282.
21. Hardin JW, Hilbe JM. *Generalized Estimating Equations.* Chapman & Hall/CRC; 2002.