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

**Corresponding Author:** Susanne G. Barnett, University of Wisconsin-Madison, School of Pharmacy, 1014 Rennebohm Hall, 777 Highland Ave., Madison, WI 53705-2222. Tel: 608-265-1168. Email: susanne.barnett@wisc.edu

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

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

**Methods.** Six schools/colleges of pharmacy each submitted ten 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 schools/colleges of pharmacy 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 considered 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 graduates be able to verbally and non-verbally communicate with individuals, groups, and organizations. Communication is also emphasized as a central part of the Pharmacists’ Patient Care Process (PPCP). 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.

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, consistent performance-based assessments and reliable evaluation tools, such as
communication rubrics, are needed. A review of the literature shows that various teaching strategies are utilized to accomplish this imperative task. There are also numerous rubrics that have been proposed to assess communication skills within various environments. To our knowledge, no study has utilized a coordinated evaluator training with a standardized codebook in an effort to increase 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. There is no rubric for assessment of communication skills among student pharmacists which 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 (use among multiple evaluator types) with standardized training to assess student pharmacist communication skills. The goal of this study was to evaluate interrater reliability of the BTAA-PBAC developed communication rubric 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. 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/colleges acted as a site for rater recruitment and training and submitted ten 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 lab 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 (1st-, 2nd-, and 3rd-year Doctor of Pharmacy students). However, all ten videos submitted by a given school/college were taken from the same simulated activity. Each skills lab 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/colleges. In the event that a school did not utilize a rater group to evaluate student performance (eg, a school did not utilize 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 pre-recorded 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/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 was not obtained.

If a rater did not pass certification, a school/college re-recruited for 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 would not be recruited and only one SP from this 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/college 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.\textsuperscript{14}

Raters returned completed paper communication rubrics to study coordinators, who entered all data into Qualtrics (Provo, UT). Gwet’s agreement coefficient was used to quantify chance-corrected agreement among several raters scoring multi-level (“no”, “inconsistent”, “yes”) items.\textsuperscript{15-16} Gwet’s agreement coefficient was chosen due to its ability to avoid paradoxes that plague kappa. 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.\textsuperscript{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 weak agreement, 0.6-0.79 moderate agreement, 0.8-0.9 strong agreement, and >0.90 almost perfect agreement.\textsuperscript{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”.\textsuperscript{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/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 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) 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 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, post-graduate year 1 resident, and student pharmacist rater groups. The SP rater group exhibited the highest agreement score of 0.77 (95% CI 0.7-0.84) 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=0.03).

DISCUSSION

This study evaluated 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.\textsuperscript{10,11} The Patient-Centered Communication Tools (PaCT) developed by Grice and colleagues demonstrated a statistically 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.\textsuperscript{10} Similarly, McKeirnan and colleagues concluded that the majority of elements on their rubric demonstrated a low level of interrater agreement.\textsuperscript{11} The current study expands upon these studies by incorporating recommendations to improve interrater reliability. Previous authors emphasize the importance of training evaluators, with Grice and colleagues specifically recommending the use of multiple videos.\textsuperscript{10-11} 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% of 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.\textsuperscript{10} 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 utilized 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 versus a detailed point-based assessment. For this
reason, the rubric may be more appropriate for use in the didactic setting compared to 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 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 nonjudgmental attitude” (item 5), “listened to and
engaged with patient” (item 6), 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 “avoided 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, using the
word “perfect” after a patient states they have diabetes, may be more effective at identifying inappropriate filler words and
make this item less subjective.

Another item with weak agreement was “provided opportunity for and responded to questions” (item 17). Of note,
among rater groups, SPs had high agreement on this item while faculty raters had very low agreement. This may represent
the tendency of faculty to assess the appropriateness of the response which was not intended to be evaluated. 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 the appropriateness of 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 but should also be a point of emphasis during the 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. It could
be argued 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 schools/colleges of pharmacy. 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 interrater 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

Table 1. Definitions of rater groups for study inclusion.

<table>
<thead>
<tr>
<th>Rater Group</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faculty</td>
<td>A professor of any ranking, or a fellow or post-graduate year-2 resident focusing on teaching in the laboratories.</td>
</tr>
<tr>
<td>Standardized patient</td>
<td>An individual acting in the role of a patient (often with a script) in an attempt to provide a standardized experience to multiple students.</td>
</tr>
<tr>
<td>Postgraduate year 1 resident</td>
<td>An individual with a Doctor of Pharmacy degree who is enrolled in a postgraduate year-1 resident program.</td>
</tr>
<tr>
<td>Student pharmacist</td>
<td>Any individual currently enrolled in a Doctor of Pharmacy program.</td>
</tr>
<tr>
<td>Preceptor</td>
<td>A pharmacist who serves as an instructor for Doctor of Pharmacy students at an experiential site.</td>
</tr>
</tbody>
</table>

Table 2. Participant recruitment, exclusion, and inclusion by rater group.

<table>
<thead>
<tr>
<th>Rater Group</th>
<th>Initially Recruited</th>
<th>Did Not Pass Certification</th>
<th>Lost to Follow-Up</th>
<th>Re-recruited and Passed Certification</th>
<th>Excluded from Analysis</th>
<th>Included in Analysis</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faculty^a</td>
<td>11</td>
<td>--</td>
<td>--</td>
<td>--</td>
<td>11</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Standardized patient^b</td>
<td>10</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>9</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>Postgraduate year 1 resident</td>
<td>12</td>
<td>--</td>
<td>1</td>
<td>1</td>
<td>12</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Student pharmacist^c</td>
<td>14</td>
<td>3</td>
<td>--</td>
<td>1</td>
<td>12</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>Preceptor^d</td>
<td>11</td>
<td>1</td>
<td>--</td>
<td>1</td>
<td>11</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>All raters</td>
<td>58</td>
<td>6</td>
<td>2</td>
<td>5</td>
<td>55</td>
<td>55</td>
<td></td>
</tr>
</tbody>
</table>

^a One college/school was able to recruit one versus two raters in the faculty and preceptor rater groups.

^b Standardized patients were utilized and recruited by five of the six participating schools/colleges.

^c One college/school initially recruited two additional student raters.
Table 3. Communication Rubric Item Rater Agreement Scores.

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

*Agreement score was calculated using Gwet’s unweighted agreement coefficient. Scores <0.39 were interpreted as no to minimal agreement, 0.4-0.59 weak agreement, 0.6-0.79 moderate agreement, 0.8-0.9 strong agreement, and >0.90 almost perfect agreement.
Table 4. Rater Group Agreement Comparisons.

<table>
<thead>
<tr>
<th>Rater Group</th>
<th>Average Agreement Score* (95% CI)</th>
<th>Deflection from Overall Average Agreement Score (95% CI)</th>
<th>Rater Group Comparison to Overall Average Agreement Score (P value)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faculty</td>
<td>0.72 (0.64, 0.80)</td>
<td>-0.01 (-0.06, 0.07)</td>
<td>0.54</td>
</tr>
<tr>
<td>Standardized patient</td>
<td>0.77 (0.70, 0.84)</td>
<td>0.03 (-0.001, 0.07)</td>
<td>0.06</td>
</tr>
<tr>
<td>Postgraduate year 1 resident</td>
<td>0.75 (0.67, 0.83)</td>
<td>0.02 (-0.02, 0.06)</td>
<td>0.33</td>
</tr>
<tr>
<td>Student pharmacist</td>
<td>0.75 (0.66, 0.84)</td>
<td>0.02 (-0.02, 0.05)</td>
<td>0.35</td>
</tr>
<tr>
<td>Preceptor</td>
<td>0.68 (0.58, 0.78)</td>
<td>-0.05 (-0.10, -0.01)</td>
<td>0.03</td>
</tr>
<tr>
<td>Overall (all groups)</td>
<td>0.73 (0.66, 0.81)</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

Abbreviations: CI = confidence interval
*Average agreement score was calculated using Gwet’s unweighted agreement coefficient over rubrics items 1-20 for each rater group and overall. The global assessment score was excluded from this analysis.

Figure 1. Communication Rubric Item Agreement Scores, All Raters.

Abbreviations: G=Global Assessment
*Agreement scores <0.39 were interpreted as no to minimal agreement, 0.4-0.59 weak agreement, 0.6-0.79 moderate agreement, 0.8-0.9 strong agreement, and >0.90 almost perfect agreement.