The Boston bowel preparation scale: a valid and reliable instrument for colonoscopy-oriented research

Edwin J. Lai, MD, Audrey H. Calderwood, MD, Gheorghe Doros, PhD, Oren K. Fix, MD, MSc, Brian C. Jacobson, MD, MPH, FASGE
Boston, Massachusetts, USA

Background: Few bowel-preparation rating scales have been validated. Most scales were intended for comparing oral purgatives and fail to account for washing and/or suctioning by the endoscopist. This limits their utility in studies of colonoscopy outcomes, such as polyp-detection rates.

Objective: To develop a valid and reliable scale for use in colonoscopy outcomes research.

Setting: Academic medical center.

Methods: We developed the Boston bowel preparation scale (BBPS), a 10-point scale that assesses bowel preparation after all cleansing maneuvers are completed by the endoscopist. We assessed interobserver and intraobserver reliability by using video footage of colonoscopies viewed on 2 separate occasions by 22 clinicians. We then applied the BBPS prospectively during screening colonoscopies and compared BBPS scores with clinically meaningful outcomes, including polyp-detection rates and procedure times.

Results: The intraclass correlation coefficient (a measure of interobserver reliability) for BBPS scores was 0.74. The weighted kappa (a measure of intraobserver reliability) for scores was 0.77 (95% CI, 0.66-0.87). During 633 screening colonoscopies, the mean (SD) BBPS score was 6.0 ± 1.6. Higher BBPS scores (≥5 vs <5) were associated with a higher polyp-detection rate (40% vs 24%, P < .02). BBPS scores were inversely correlated with colonoscope insertion (r = -0.16, P < .003) and withdrawal (r = -0.23, P < .001) times.

Limitations: Single-center study.

Conclusions: The BBPS is a valid and reliable measure of bowel preparation. It may be well suited to colonoscopy outcomes research because it reflects the colon’s cleanliness during the inspection phase of the procedure. (Gastrointest Endosc 2009;69:620-5.)
Furthermore, in any individual patient, the quality of bowel preparation may vary between colonic segments. It might prove useful to have a bowel-preparation rating scale that is sensitive to such differences to better define the likelihood of a missed polyp and/or appropriate screening and surveillance intervals. We sought to develop a novel bowel-preparation rating scale specifically for application during withdrawal of the colonoscope, after all cleansing maneuvers are completed. Such a scale could be used in the clinical and research settings, controlling for bowel preparation in studies that assess rates of missed lesions and for establishing guidelines on appropriate screening and surveillance intervals inclusive of bowel preparation quality.

PATIENTS AND METHODS

The study was approved by the institutional review board of the Boston University Medical Center.

Development of the Boston bowel preparation scale

The Boston bowel preparation scale (BBPS [suggested pronunciation “bee-bops”]) was developed to limit interobserver variability in the rating of bowel-preparation quality, while preserving the ability to distinguish various degrees of bowel cleanliness. Subjective terms, such as “excellent,” “good,” “fair,” “poor,” and “unsatisfactory,” are replaced by a 4-point scoring system applied to each of the 3 broad regions of the colon: the right side of the colon (including the cecum and ascending colon), the transverse section of the colon (including the hepatic and splenic flexures), and the left side of the colon (including the descending colon, sigmoid colon, and rectum). The points are assigned as follows (Fig. 1):

- 0, unprepared colon segment with mucosa not seen because of solid stool that cannot be cleared.
- 1, portion of mucosa of the colon segment seen, but other areas of the colon segment are not well seen because of staining, residual stool, and/or opaque liquid.
- 2, minor amount of residual staining, small fragments of stool, and/or opaque liquid, but mucosa of colon segment is seen well.
- 3, entire mucosa of colon segment seen well, with no residual staining, small fragments of stool, or opaque liquid.

The wording of the scale was finalized after incorporating feedback from 3 colleagues experienced in colonoscopy.

Each region of the colon receives a “segment score” from 0 to 3, and these segment scores are summed for a total BBPS score, which ranges from 0 to 9. Therefore, the maximum BBPS score for a perfectly clean colon without any residual liquid is 9, and the minimum BBPS score for an unprepared colon is 0. If an endoscopist aborts a procedure because of inadequate preparation, then any nonvisualized proximal segments are assigned a score of 0.

Assessment of reliability

To enhance comprehension of the BBPS, a 15-minute training digital video disc (DVD) was created and viewed by members of our gastroenterology division. The DVD contains narrated video footage that illustrates each point of the BBPS. It also illustrates how a segment score may be improved through maneuvers such as washing and fluid aspiration. Next, 2 truncated demonstration colonoscopies (only the withdrawal portion) were included to show how the BBPS would be applied; these demonstration colonoscopies exhibited total BBPS scores considered to be 4 and 5, respectively. Copies of the BBPS DVD may be obtained from the corresponding author (see reprint requests line at the end of this article).

Assessment of validity

The training DVD also contains 3 truncated testing colonoscopies, with images that differ from those in the demonstration colonoscopies. The 3 testing colonoscopies had bowel-preparation qualities considered to represent total BBPS scores 4, 5, and 6, respectively. To assess reliability, we asked members of our gastroenterology division to rate the quality of bowel preparation in each testing colonoscopy by using the BBPS. Participants viewed the testing colonoscopies on 2 occasions, at least 1 month apart. For the second viewing, the order of the testing colonoscopies was changed to limit the possibility that someone might remember the scores that he or she had provided during the first viewing. The scores from the 2 viewings were used to calculate intraobserver and interobserver reliability.

Capsule Summary

What is already known on this topic

- Published bowel-rating scales measure the degree of bowel cleanliness encountered by endoscopists during initial inspection of the colon.

What this study adds to our knowledge

- The Boston bowel preparation scale, a 10-point scale that assesses bowel preparation after the endoscopist’s cleansing maneuvers, demonstrated good intraobserver and interobserver reliability and was favorably associated with clinical outcomes, such as polyp-detection rates, recommendations for repeated procedures, and colonoscope insertion and withdrawal times.
colonoscopy, the endoscopist was asked to record the quality of bowel preparation by using both the categorical system used historically at our medical center ("excellent," "good," "fair," "poor," or "unsatisfactory") and the BBPS score. The endoscopists also recorded the location and size of all polyps found during the examination, as well as whether they were recommending a repeated colonoscopy specifically because the bowel preparation was deemed inadequate. Endoscopy nurses recorded colonoscope insertion and withdrawal times.

To measure the construct validity of the BBPS, we assessed 4 factors: (1) comparison with another, albeit nonstandardized, method of assessing bowel preparation (ie, excellent, good, fair, poor, unsatisfactory), (2) the association between the BBPS score and a perception of inadequate bowel preparation, (3) the association between the BBPS score and the polyp-detection rate, and (4) the association between the BBPS score and colonoscope insertion and withdrawal times.

**Statistical analysis**

To assess interobserver reliability, we calculated the intraclass correlation coefficient (ICC) among scores applied after viewing the testing colonoscopies. Because, for each testing case, 2 BBPS scores were available from each clinician, we randomly selected 1 of the 2 scores for this calculation. We repeated this procedure 1000 times to determine the potential distribution and a 95% predictive interval for the ICC of a single reading. To assess intraobserver reliability, we calculated weighted kappa measures. We calculated the mean total BBPS score for each possible categorical assessment: "excellent," "good," "fair," "poor," and "unsatisfactory," and obtained a P value for the trend in means by using linear regression. We determined the polyp-detection rate for each BBPS score, as well as for a dichotomized score (<5 and ≥5). This dichotomized point was chosen a priori based on a clinical assessment that the degree of cleanliness that causes a score <5 would likely be considered inadequate. Associations between BBPS scores and polyp-detection rates, as well as recommendations for repeated procedures, were calculated by using χ² tests. Colonoscope insertion and withdrawal times were correlated with BBPS scores by using the Pearson correlation coefficient. For colonoscope withdrawal times, we excluded cases in which polyps were found. All calculations were performed by using SAS version 9.2 (SAS Institute, Cary, NC), and 2-sided P values <.05 were considered significant.

**Figure 1.** The BBPS. A, Segment score 0, unprepared colon segment with mucosa not seen because of solid stool that cannot be cleared. B, Segment score 1, portion of mucosa of the colon segment seen, but other areas of the colon segment not well seen because of staining, residual stool, and/or opaque liquid. C, Segment score 2, minor amount of residual staining, small fragments of stool and/or opaque liquid, but mucosa of colon segment seen well. D, Segment score 3, entire mucosa of colon segment seen well with no residual staining, small fragments of stool and/or opaque liquid.
RESULTS

The BBPS training and testing DVD was viewed by 22 members of our gastroenterology section, including 13 full-time faculty, 8 fellows, and 1 physician assistant with more than 10 years’ experience in performing flexible sigmoidoscopy. Individuals viewed the DVD twice, with a mean (SD) of 10 ± 3 weeks between viewings. The ICC for interobserver agreement of a single reading in total BBPS scores was 0.74 (95% predictive interval 0.67-0.80). The weighted kappa value for intraobserver agreement in total BBPS scores was 0.77 (95% CI, 0.66-0.87). This degree of agreement is considered to be substantial.10 The ICCs and weighted kappa values stratified by experience (ie, attendings vs fellows) are shown in Table 1.

When the BBPS was used prospectively during 633 screening colonoscopies, we observed an approximate bell-shaped distribution of scores (Fig. 2). The mean (SD) BBPS score was 6.2 ± 1.5, and the median score was 6.0 (range 0.0-9.0; interquartile range 6.0-7.0). When considering the categorical bowel-preparation ratings used during those colonoscopies (excellent, good, fair, poor, and unsatisfactory), we noted a significant decreasing trend in the mean BBPS score assigned in each category (P for trend <.001) (Fig. 3).

Among the 633 patients who underwent colonoscopy, 243 had at least 1 polyp found (38%). The polyp-detection rate for each BBPS score was as follows: 0, 0%; 1, 0%; 2, 33%; 3, 19%; 4, 33%; 5, 43%; 6, 45%; 7, 31%; 8, 35%; and 9, 36%. The polyp-detection rate was 40% for patients with a BBPS score ≥5 compared with 24% for patients with a BBPS score <5 (P < .02). The endoscopist recommended repeating the procedure because of inadequate bowel preparation among 2% of cases with a BBPS score ≥5 compared with 73% among cases with a BBPS score <5 (P < .001). Total BBPS scores were inversely correlated with both colonoscopy insertion times (r = -0.16, P < .003) and withdrawal times (r = -0.23, P < .001).

DISCUSSION

We developed a valid and reliable bowel-preparation rating scale that can be easily taught with a brief instructional DVD. The BBPS demonstrated good intraobserver and interobserver reliability among 22 physicians, including both fellows and attendings. Prospective use of the BBPS during screening colonoscopy showed significant associations with clinical outcomes, such as polyp-detection rates, recommendations for repeated procedures, and colonoscope insertion and withdrawal times.

Many previously published bowel-rating scales were designed specifically to compare the efficacy of 2 or more bowel-preparation methods.11-15 As such, they measure the degree of bowel cleanliness encountered by endoscopists during initial inspection of the colon. The BBPS distinguishes itself from these scales by being applied after the endoscopist has performed any additional cleansing maneuvers, which reflects the actual practice of colonoscopy. Therefore, the BBPS may be better suited to colonoscopy-outcomes research, such as studies aimed at defining appropriate screening and surveillance intervals that account for bowel-preparation quality. Furthermore, the BBPS can also be used when comparing bowel preparations. In such instances, the study outcome would represent the clinical effectiveness of the preparations tested (eg, “Did Mrs Jones have better colonoscopic visualization after using preparation A versus B?”) instead of the efficacy of the preparations (ie, Does one preparation clean better than the other?). This is an important distinction, because, without accounting for an endoscopist’s ability to improve preparation quality with cleansing maneuvers during colonoscopy, the clinical impact of one preparation versus another remains unknown.

Many published bowel-preparation scales rely on a global assessment of bowel cleanliness, which fails to account for differences in individual colon segments. During colonoscopy, however, one may find a generally excellent preparation, except for 1 region that is poorly prepared. The BBPS recognizes that the colon is not uniformly

<table>
<thead>
<tr>
<th>TABLE 1. Reliability of the BBPS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>---------------------------------</td>
</tr>
<tr>
<td>Interobserver reliability</td>
</tr>
<tr>
<td>No. participants</td>
</tr>
<tr>
<td>Intraclass correlation coefficient</td>
</tr>
<tr>
<td>Intraobserver (test-retest) reliability</td>
</tr>
<tr>
<td>No. participants</td>
</tr>
<tr>
<td>Weighted kappa (95% CI)</td>
</tr>
</tbody>
</table>

*Includes 1 physician assistant with more than 10 years of experience performing flexible sigmoidoscopy.
| One physician was unable to view the DVD twice. |
prepared for colonoscopy, which allows the assignment of various scores to each of 3 broad segments of the colon. By accounting for such subtleties, the BBPS may help to better define risks for missed pathology, although this remains to be demonstrated. Other published bowel-preparation scales rely on factors prone to interobserver variation, such as quantitative estimates of residual stool or liquid, the percentage of visualized mucosa, or the likelihood of missing certain sized lesions. The BBPS relies on more generalized assessments by using segment scores to permit tailoring to individual patients.

Few of the previously published bowel-preparation rating scales have been formally validated. The Aronchick scale was evaluated by 5 gastroenterologists who reviewed 80 videotaped colonoscopies. Interobserver reliability was measured by using ICCs that ranged from 0.31 for “distal colon to hepatic flexure” to 0.76 for the cecum. A Friedman \( \chi^2 \) test was also used to test the likelihood that samples of given scores were drawn from the same population. Intraobserver reliability was not reported, nor was there formal correlation with other colonoscopy outcomes such as polyp-detection rates.

Another validated scale, the Ottawa bowel preparation scale, uses 3 colonic segment scores (in this case 0-4) that are summed as part of a total score. However, there is an additional global fluid-quantity rating (0-2), which requires subjective estimation of residual liquid. The Ottawa bowel preparation scale was validated only by comparison with the Aronchick scale and not by correlation with colonoscopy outcomes. Reliability testing was limited to 2 observers, a staff gastroenterologist and a research fellow, who observed 97 colonoscopies. Intraobserver reliability was tested by using Pearson correlation coefficients, linear regression analyses, and a kappa ICC. This scale performed well, albeit between only 2 investigators, with a kappa ICC of 0.94 (95% CI, 0.91-0.96), but intraobserver reliability was not assessed.

We believe that the BBPS has now been reasonably validated for general use in research studies. However, our reliability testing was based on 3 truncated colonoscopy video clips that reflected BBPS scores in the mid scale range (considered to be 4, 5, and 6), rather than full colonoscopies that reflect all 9 BBPS scores. We chose to test the reliability in the mid scale range, postulating that this would be the region with the broadest interobserver variability. Moreover, we postulated a priori that the clinically relevant cut point regarding a preparation’s overall adequacy would likely fall in this range. In addition, there is likely very good agreement between gastroenterologists assessing excellent and poor preparations, but this will need to be proven in future studies.

The strengths of our study include the large number of individuals who participated in reliability testing and the large number of cases and clinically meaningful outcomes used to prospectively validate the scale. However, our study was limited to a single institution, potentially limiting the generalizability of our results. It is reassuring that we found similar results among fellows, attendings, and a GI physician assistant, which suggests that the BBPS can be used by clinicians with various levels of experience. Furthermore, the BBPS training DVD is brief (15 minutes, including testing videos), which makes dissemination of the scale, and standardization of its use, straightforward.

Unfortunately, we are unable to comment on the utility of the BBPS during other procedures that require colonic catharsis, eg, CT colonography. It is not clear that the BBPS can be used effectively in noncolonoscopy bowel imaging, particularly because the distinction between segment

![Figure 2](image-url) Distribution of BBPS scores applied during 633 screening colonoscopies. The distribution approximates a bell-shaped curve, with a median score of 6 and an interquartile range of 6-7.

![Figure 3](image-url) We determined the mean (SD) BBPS score (y-axis) assigned for each categorical bowel preparation rating (x-axis) observed during 633 screening colonoscopies. The decreasing trend in scores across decreasing categorical preparation qualities was statistically significant (P < .001). Categorical bowel preparation rating: 1 = excellent, 2 = good, 3 = fair, 4 = poor, 5 = unsatisfactory.
scores 2 and 3 is likely impossible without direct visualization of the bowel. Furthermore, we did not measure the reliability of the BBPS in noncolonoscopy settings.

In summary, the BBPS is a valid and reliable instrument for rating the quality of bowel preparation during colonoscopy. Investigators may find it useful for colonoscopy-oriented research that requires a method of controlling for various degrees of bowel preparation. Future studies should assess the validity of the BBPS at other institutions, verify its reliability across the full spectrum of scores, and examine the relationship between individual segment scores and polyp-detection rates.

REFERENCES


Received April 11, 2008. Accepted May 18, 2008.

Current affiliations: Section of Gastroenterology, Boston University Medical Center (E.J.L., A.H.C., B.C.J.), the Department of Biostatistics, Boston University School of Public Health (G.D.), Boston, Massachusetts, the Division of Gastroenterology, University of Washington (O.K.F.), Seattle, Washington, USA.


Reprint requests: Brian C. Jacobson, MD, Section of Gastroenterology, Boston University Medical Center, 85 East Concord St, Rm 7721, Boston, MA 02118.

If you want to chat with an author of this article, you may contact him at brian.jacobson@bmc.org.