Comparison of Two Methods Used for Final Internship Evaluation at St. Marianna University Hospital

Atsushi Mochizuki¹, Hiroko Tina Tajima¹, Takayuki Asano², and Miyuki Ino¹

(Received for Publication: September 16, 2020)

Abstract

During the first stage of clinical internship at St. Marianna University Hospital, we believe that it is important to conscientiously consider the “process of reflection” and consolidate the growth trajectory and outcomes. This led to the introduction of portfolios in 2003 that considers the medical interns and their instruction to be a single unit, allows them to establish objectives, and facilitates their reflection of their own behavior. Our subjects were 43 interns who started the internship program at St. Marianna University Hospital in 2012 and underwent the Final Internship Evaluation after completing the first stage of internship in March 2014. The new Final Internship Evaluation criteria (rubric) included six evaluation perspectives. An evaluation scale was created to grade each question on a scale of 1 to 5. In addition, the valuation was also performed using the conventional method, which included 15 criteria, to investigate correlation between the portfolio evaluation method and the conventional evaluation method. Five points could be scored in total for a single item, thus allowing interns to attain a total score of 75 points. The mean evaluation score given by the two instructors using the new evaluation method was 4.4 points, whereas that given using the conventional method was 57.1 points. The performance rankings obtained with the old and new evaluation methods were similar, and a positive correlation was observed between the two (correlation coefficient: 0.69). The grades awarded after using the new and old evaluation methods were similar even though the evaluation perspectives in these two methods differed, so it would be desirable to perform an evaluation that combines elements of both methods.

Key Words
Portfolio, rubric, reflection, internship, evaluation

Introduction

There is a need for evaluation of medical interns during their clinical training, after graduation, and such evaluation involves subdividing the required learning into predefined learning objectives, selecting evaluation methods with evidence of high validity and reliability, and using these methods to evaluate medical interns. We believe that during the first stage of clinical internship at St. Marianna University Hospital, it is important to conscientiously consider the “process of reflection” and consolidate the growth trajectory and outcomes. This notion led to the introduction, in 2003, of learning portfolios, which consider the medical intern and instructor to be a single unit, allow them to establish objectives, and facilitate self-reflection in terms of the intern’s own behavior. These learning portfolios are practically applied during formative and comprehensive evaluations.

Between 2003 and 2011, the Final Internship Evaluation, which was conducted at the end of two years of early clinical training, was a comprehensive evaluation determined by the total number of points attained in a five-step evaluation covering a total of 15 items and entitled the Primary Clinical Internship Portfolio Evaluation. This evaluation also covered the attainment of objectives (general objectives, behavioral objectives, experience objectives) required by the

¹ Research Institute for Medical Education, St. Marianna University School of Medicine
² Division of Gastroenterological and General Surgery, St. Marianna University School of Medicine
Ministry of Health, Labor and Welfare and involved submission of various documents (employment status documents and proof of participation in educational programs, among others). The 15 items included four on composition, five on contents, and six on overall evaluation. A total of 5 points could be scored for a single item, making it possible for interns to attain a total score of 75 points. An intern’s portfolio was read by two clinical internship instructors who scored the portfolio separately, and the final score was the average of the points awarded by these two instructors.

The intern’s portfolio is used to assess the practical training that the intern has undertaken during the two-year period as well as the intern’s reflection on this training. However, the fact that interns have numerous interviews with instructors over time and that they read a great deal of material allows not only the interns but also the instructors to obtain feedback that indicates whether or not the instructive activities are appropriate. We believe that this allows both interns and instructors to improve over time. However, 10 years have passed since introduction of the portfolio system, and when we reviewed the evaluation process, we found that the 15-item Final Internship Evaluation may be assessing only the degree of completeness of the deliverable known as the portfolio. We therefore conducted a review of other evaluation methods that would allow for a performance-based evaluation of an intern’s knowledge, ability, and attitude. For this reason, and to ensure objectivity during comprehensive evaluation, we established new criteria (now known as the rubric) by which to evaluate the intern’s achievements \(^1\),\(^2\), and we introduced these criteria during the 2012 internship program. In February 2014, when the interns in the 2012 program completed their internship, we used the rubric to perform comprehensive evaluation for the Final Internship Evaluation, and we compared this method to the previous evaluation method.

Subjects and Methods

Our subjects were 43 interns who started the internship program at St. Marianna University Hospital in 2012 and were subjected to the Final Internship Evaluation after completing the first stage of clinical internship in March 2014. The interns were shown the criteria (rubric) for the Final Internship Evaluation in advance, either before they started the program or during the orientation when they started employment (Table 1). The new Final Internship Evaluation criteria (rubric) included six evaluation perspectives: (1) employment status/document submission status, (2) creation status of the internship records, (3) attainment status of basic skills, (4) acquisition status of comprehensive knowledge, and (6) attitude during the internship (Table 1). An evaluation scale was created so that each item could be graded on a scale of 1 to 5, representing five possible levels of achievement. The evaluation was performed by two clinical internship instructors nominated by the clinical internship center, and each placed a mark next to each item and considered the highest level to be the attainment level from that evaluation perspective. A final score was obtained by calculating the mean value of the points awarded by each of the two instructors. However, if a candidate was awarded a level 1 (inadequate) score for the first evaluation perspective, the documents were reconsidered after corrections were made in conjunction with the supervising instructor, and if the candidate was ultimately awarded a level 1 score without correction, the candidate was not able to complete the internship. In addition, the same 43 candidates were evaluated by means of the conventional method so that correlation between the new method and the conventional evaluation method could be investigated (Table 2). The conventional evaluation was performed using the 15 criteria proposed by Driessen et al.\(^3\), and it included four items on portfolio composition, five items on portfolio contents, and six items on overall evaluation. A maximum of 5 points (Likert score) could be given for a single item, so it was possible for an intern to attain a total score of 75 (15 items x 5) points.

This study was approved by St. Marianna University School of Medicine Ethics Committee (approval no.: 4925).

Results

The mean evaluation score given by the two instructors using the new evaluation method was 4.4 (2.0 to 5.0, 88.0%) points, and none of the candidates was ultimately awarded an “inadequate” (a level 1 score). Meanwhile, the mean evaluation score given by the two instructors using the conventional method was 57.1 (34.5 to 74.0, 76.1%) points (Fig. 1).

In addition, we performed a comparison to determine whether the scores awarded to an intern by means of the two evaluation systems represent the same or different levels of achievement. We did this by using the two methods simultaneously when performing the Final Internship Evaluation for interns.
Table 1. Internship Evaluation Rubric (partial)

| Ability to gather data | Diagnostic skills | Self-analysis (Reflection) |
|------------------------|-------------------|---------------------------|
| **5** (Excellent)      | The documents submitted were all discussed with and signed by the instructor by the end of the rotation and provide evidence of internship activities. | A Mini-CEX was performed in each treating department, and evidence showing evaluation of simulations and OSCEs has been submitted. | During significant event analysis (SEA), the intern has evaluated and analyzed how they feel and mentioned future challenges. |
| **4** (Very good)      | Almost all documents submitted were discussed with and signed by the instructor by the end of the rotation and provide evidence of internship activities. | A Mini-CEX was performed in some of the treating departments, and evidence showing evaluation of simulations and OSCEs has been submitted. | The details mentioned above have not been explored in detail, and it is written in the intern’s own words, but the expression lacks clarity. |
| **3** (Good)           | The documents submitted were all discussed with and signed by the instructor by the submission deadline and provide evidence of internship activities. | A Mini-CEX was performed in some of the treating departments, and evidence showing evaluation of simulations and OSCEs has been submitted, although the evidence is insufficient. | The objectives attained, points for improvement, feelings and emotions, and details and desires regarding future learning that form part of the self-evaluation were captured to a certain extent, but not completely. |
| **2** (Poor)           | The evaluations of the training period were collected, but the content was biased in some areas. | There was a lack of evidence to show that simulation, OSCE, and other procedures in each department were evaluated. | The self-assessment included a description of goals achieved, things to improve on, feelings, emotions, content and aspirations to learn in the future, but the description was not enough. |
| **1** (Inadequate)     | The collection of evaluations during the training period was completely inadequate. | There was no evidence that they were committed to mastering medical techniques during all training periods. Mini-CEX was not performed as an assessment of medical procedures. | The self-assessment either did not describe the goals achieved, things to improve on, feelings, emotions, content and aspirations to learn in the future, or even if it did, the content was inadequate. |

scheduled to complete the program in 2012. Use of the conventional evaluation method yielded a mean score of 57.1 (34.5 to 74.0) points (out of a possible 75 points), and four of the five interns ranked at the top by the conventional evaluation method achieved level 5 scores by means of the new method (one intern scored 4.5). The rankings obtained by means of both evaluation methods were similar, and positive correlation was observed between scores conferred by means of the two methods (correlation coefficient: 0.69, Fig. 2).

Discussion

The items covered by the new and old evaluation methods differ, since the old method evaluates the degree of completeness of the portfolio (composition, contents, etc.), whereas the new method incorporates evaluation of various competencies\(^4\). Skills such as self-management and self-evaluation are elements of lifelong learning\(^5\) and should always be subject to formative evaluation\(^7\).

Going forward, one of the challenges will be the
Table 2. Evaluation Sheet for the Conventional Method

1. Was the layout above average?
2. Was the sentence construction above average?
3. Was the composition above average?
4. Was the portfolio complete?
5. Was the intern able to perform critical (objective) observation?
6. Was an external and internal explanation included in the self-analysis of the aforementioned strengths and weaknesses? Was the self-analysis only limited to a list of facts and circumstances?
7. Has the aforementioned analysis of strengths and weaknesses been conducted from an adequately broad perspective including different roles in each of the departments during the internship rotations?
8. Does the portfolio consistently exhibit items that can be used as evidence?
9. Did the intern become aware of an increase in their own abilities in the portfolio?
10. In terms of the portfolio contents, did the intern achieve more than the learning objectives in the internship program?
11. Did the intern make practical application of portfolios that were created previously?
12. Was there a logical organization based on the strengths and weaknesses of the internship role, and was the intern able to clearly define their learning objectives?
13. Does the portfolio mention whether the intern asked to what extent the learning objectives were achieved?
14. Did the intern exert effort to attempt to achieve the attainment of objectives in each of the internship rotations?
15. Do you feel that the intern spent an appropriate amount of time on this portfolio?

Five-stage evaluation: Evaluations range from 1: completely inappropriate to 5: excellent, and the intern may obtain a total of 75 points

![Rubric](chart1.png)  
![Conventional method](chart2.png)

**Figure 1.** Mean value of the evaluation results after using the conventional method and the rubric

The fact that it is impossible to completely eliminate inter-evaluator error solely through use of the rubric. For this reason, we need to continue to hold workshops in order to control inter-evaluator error. It is difficult to create levels between the criteria when creating the rubric, so we need to further improve the level classification for each item to ensure validity of the contents.

We believe that the rubric is useful for clinical internships because the old method allowed the different evaluators to apply different criteria during the evaluation, which in turn had different implications.
for the five-stage evaluation of each evaluation item. In contrast, the new evaluation method stipulates an evaluation scale for each evaluation perspective, and it allows the interns and instructors to take joint ownership of the evaluation criteria.

After assessment using the new evaluation method, we reflected on the method again, and we found that certain areas were in need of improvement. This included the fact that the rubric that was newly created included detailed numerical targets to create a detailed division of levels for each evaluation item. There were some evaluation scales (e.g., employment status) that were difficult to divide into five-stage levels, and these scales were unreasonably divided into five stages in order to combine them with other evaluation items. We believe that this led to an evaluation table that was difficult for both interns and evaluators to understand.

The newly established internship evaluation criteria (rubric) is made up of six evaluation perspectives that are subject to a five-stage evaluation (5: excellent, 4: very good, 3: good, 2: requires further effort, 1: inadequate [recreate]). However, the criteria for a level 3 score and the criteria for a level 4 score are similar, and we believe that the two levels can be combined. In addition, complex evaluation criteria are mentioned in the evaluation scale for a single evaluation perspective, and this may be puzzling to both interns and evaluators. We plan to improve the method by first converting the current five-stage scale into a four-stage scale (4: highly successful [highest level], 3: successful [standard level], 2: requires improvement, 1: inadequate [recreate]), which will allow an adjustment to ensure the combination of a single evaluation criteria for a single item in an evaluation perspective. In addition, there should be clarification of the evaluation procedure for evaluation criteria in terms of psychomotor performance (ability), which is difficult to evaluate in the portfolio\(^8\). Furthermore, the weighting (evaluation ratio) of each evaluation perspective should be clarified, and the specific weight for each evaluation perspective should emphasize the relative degree of importance of each aspect (Table 1). This is not to say that inter-evaluator error will be eliminated purely by joint ownership of the rubric between interns and instructors, so we need to hold periodic workshops aimed at minimizing such error\(^9\).

By decreasing the number of evaluation scales, the difference between each stage will become clearer and will allow an adjustment to ensure the combination of a single evaluation criteria for a single item in an evaluation perspective, and we expect that this will clarify the reason why interns have received a particular score\(^8\). We believe that clarification of the evaluation procedures for each evaluation criteria in an evaluation perspective will also clarify what kind of internship records\(^10,11\) and evidence will be used as the basis for the evaluation\(^8\). Showing the weight (evaluation ratio) of each evaluation perspec-

---

**Figure 2.** Correlation between the conventional method and the rubric
Correlation coefficient: 0.69
tive does not simply equalize the evaluation of employment status records and self-analysis ability or attitudes during the internship. However, we believe that it will once again communicate the relative importance of the abilities that should be acquired by the end of internship during the Final Internship Evaluation9).

Holding periodic workshops to minimize the number of inter-evaluator errors is important for increasing the reliability of the evaluation, although periodic inspections will function as plan-do-check-act cycles at the clinical internship center, and will be expected to improve instruction methods and be associated with continuous improvement of the internship evaluation process12.

We will still need to fulfill our responsibility to explain the portfolio-based evaluation method to society. Criteria to guarantee the quality of the portfolios, the portfolio evaluation procedure, and publishing the evaluation criteria for the Final Internship Evaluation, i.e., the rubric, will do a great deal to ensure our accountability to society.

Conclusion

We found the grades awarded by means of the new and old evaluation methods to be similar, although the evaluation perspectives of these two methods may differ, so it would be desirable to perform an evaluation that combines elements of both. In addition, we believe that we need to further investigate the details of the rubric, since there are items that are difficult to divide into specific levels of achievement. Doing so will ensure the validity of each level of achievement awarded when the rubric is used for the Final Internship Evaluation.

Conflicts of Interest

The authors have no conflicts of interest to disclose.

References

1) Dawson P. Assessment rubrics: towards clearer and more replicable design, research and practice. Assessment & Evaluation in Higher Education 2017; 42: 347–360.
2) Fraile J, Panadero E, Pardo R. Co-creating rubrics: The effects on self-regulated learning, self-efficacy and performance of establishing assessment criteria with students. Studies in Educational Evaluation 2017; 53: 69–76.
3) Driessen EW, Overeem K, Tartwijk JV, et al. Validity of portfolio assessment: which qualities determine ratings? Medical Education 2006; 40: 862–866.
4) Miller GE. The assessment of clinical skills/competence/performance. Academic Medicine 1990; 65: S63–S67.
5) Panadero E, Alonso-Tapia J. Self-assessment: Theoretical and Practical Connotations. When it Happens, How is it Acquired and what to do to Develop it in our Students. Electronic Journal of Research in Educational Psychology 2013; 11: 551–576.
6) Boud D, Falchikov N. Quantitative studies of student self-assessment in higher-education: A critical analysis of findings. Higher Education 1989; 18: 529–549.
7) Jonsson A, Svingby G. The use of scoring rubrics: Reliability, validity and educational consequences. Educational Research Review 2007; 2: 130–144.
8) Oki H. How to Introduce Rubrics into Japanese Universities: Aiming at Fairness, Objectivity and Rigor in Assessment of Performance. Ritsumeikan Higher Educational Studies 2014; 14: 71–90.
9) Suzuki M. Effects of a Rubric: Values of a Test, Motivation for Learning, and Learning Strategies. The Japanese journal of educational psychology 2011; 59: 131–143.
10) Tagawa M, Saiki T. Assessment in Medical Education I. Basics. Medical Education (Japan) 2013; 44: 345–357.
11) Nishigori H, Saiki T. Assessment in Medical Education II: Detailed exposition. Medical Education (Japan) 2013; 44: 429–438.
12) Stevens DD, Levi AJ. Introduction to Rubrics, Second Edition, Stylus Publishing, LLC, 2013.