Development and Testing of a Portfolio Evaluation Scoring Tool

Karen A. Karlowicz, EdD, RN

ABSTRACT

This study focused on development of a portfolio evaluation tool to guide the assignment of valid and reliable scores. Tool development was facilitated by a literature review, guidance of a faculty committee, and validation by content experts. Testing involved a faculty team that evaluated 60 portfolios. Calculation of interrater reliability and a paired-samples t test were used to judge effectiveness. Interrater reliability was 0.78 for overall scores, 0.81 for the seven program outcomes criteria scores, and more than 0.65 for scores assigned by 11 of 13 pairs of raters. There were no significant differences between raters' scores in 10 of 13 pairs. The portfolio evaluation tool demonstrated high reliability and should be tested by other schools using portfolio evaluation.

Nursing education programs are challenged to document evidence of student achievement and competence and use that data to change or improve the curriculum (Jantzi & Austin, 2005; Tanner, 2001). Although a number of strategies are used to gauge student learning, many schools are choosing to implement portfolio evaluation as a means of judging the attainment of program outcomes (Jantzi & Austin, 2005; Karlowicz, 2000; McCready, 2007). A portfolio is a purposeful collection of assorted work assembled by the student that highlights his or her efforts, progress, and overall achievements while enrolled in a nursing course or program (Karlowicz, 2000).

Evaluation of learning through portfolio analysis gives faculty the opportunity to ascertain the acquisition of nursing knowledge and skills that may not be demonstrated by other traditional evaluation methods (Kear & Bear, 2007). For example, standardized objective testing is often used to gauge students' readiness for the NCLEX-RN®. However, this evaluation method focuses only on the demonstration of knowledge and does not provide faculty with evidence of the critical thinking processes and resources used by the students to solve clinical problems. The inherent authenticity of materials contained in portfolios is considered not only a reflection of the quality of the curriculum, but also indicators of the growth, development, and overall progress of students in terms of their preparedness for nursing practice (Wenzel, Briggs, & Puryear, 1998). Although portfolio analysis appears to have many benefits, uncertainty persists regarding its value for summative evaluation because the validity and reliability of scores have been difficult to establish. In addition, the lack of a systematic process for evaluating portfolios and assigning scores makes it difficult for nursing programs to report aggregate data regarding defined curricular outcomes.

The purpose of this study was to develop an evaluation tool that would facilitate the comprehensive appraisal of content in baccalaureate nursing student portfolios, and thus the assignment of valid and reliable scores. Valid scores are defined as ratings that reflect the presence of characteristics for each outcome represented by entries in the portfolio. Reliable scores are defined as ratings for each outcome that are consistent between two faculty evaluators.
BACKGROUND

Portfolio Format and Reliability

Regardless of academic discipline, faculty are challenged to create portfolio evaluation tools (PETs) that are easy to understand and implement, promote objective analysis, and result in scores that provide a meaningful evaluation of student performance specific to the purposes of the portfolio. The literature includes several articles that describe the experiences of faculty attempting to implement portfolios for course or program evaluation within nursing (Cook, Kase, Middelton, & Monsen, 2003; Jasper & Fulton, 2005; Kear & Bear, 2007; Martin, Kinnick, Hummel, Clukey, & Baird, 1997; Ruholl, 2000; Wenzel et al., 1998) and in other academic disciplines (Annis & Jones, 1995; Burch, 1997; Burns & Haight, 2005; MacDonald, 1997; Morgan, 1999). These studies usually focus on the content of portfolios; rarely are the processes used to judge the quality of materials contained in portfolios explained in any detail. When described, the process for portfolio evaluation is typically an explanation of a scoring rubric. A scoring rubric is a defined classification scheme to guide score assignment that is unique to the purposes of the portfolio and lists observable performance criteria, indicators, or expectations (Johnson & Rose, 1997; Taggart, Phiifer, Nixon, & Wood, 1998). The criteria and format of scoring rubrics, methods used to generate scores, and the outcomes being evaluated vary greatly. In addition, the validity and reliability of the data collected are usually not addressed.

Outcomes and competency evaluations using portfolios rely on the determination of interrater reliability to assure consistency in the independent assessment of student performance. This measure is considered standard and the basis for all other decisions about portfolio quality (Herman & Winters, 1994; Mabry, 1999). The difficulty in establishing reliability in portfolio evaluation stems from the fact that it is largely a qualitative process trying to produce quantitative results. Thus, vague terminology in the definition of scores, scoring rubrics that are too complex for most faculty to apply, a lack of understanding of the levels of student performance expected, as well as a lack of agreement regarding characteristics most valued by the evaluators can jeopardize the interpretation of evidence in portfolios and the assignment of scores (Karlowicz, 2000).

Early studies of portfolio reliability are found in the elementary education literature. Reckase (1995) explored the psychometric qualities of portfolio evaluation and determined that score reliability was comparable whether individual item or holistic scoring was used by raters. In addition, he found that the achievement of a reliability approximating 0.80 required at least 8 to 10 portfolio entries and an internal consistency measure of 0.55 for single entries and a correlation between entries of 0.28. A subsequent study (Supovitz, MacGowan, & Slattery, 1997) that examined the interrater reliability for language arts portfolios evaluated by primary grade teachers found that reliabilities between 0.58 and 0.77 resulted when there was an inconsistent number of pieces of evidence in the portfolios and raters’ interpretations of the quality of the evidence varied. Higher rates of interrater reliability for portfolio evaluation, ranging from 0.80 to 0.96, have also been reported (Naizer, 1997; Valencia & Au, 1997). The high level of consistency among portfolio evaluators in these studies is attributed to portfolio development guidelines that are clear and unambiguous, an evaluation process that is systematic, and ongoing professional development that enabled teachers to gain a better understanding of the product and process of portfolio evaluation through practice portfolio reviews.

In the area of health care education, Pitts, Coles, Thomas, and Smith (2002) conducted a study to examine the interrater reliability of assessors evaluating course portfolios by prospective general practice medical trainers in the United Kingdom. Twelve portfolios were rated individually and in teams of two by eight assessors. The study results suggested that the most appropriate way to measure interrater reliability in portfolio evaluation is to compare scores by pairs of assessors (kappa scores range from 0.01 to 0.65) rather than by individual raters (kappa scores range from 0.05 to 0.36). More recently, Kear and Bear (2007) reported a kappa of greater than 0.80 for the majority of the 22 individual items on their PET, and a Pearson’s correlation of 0.95 for total scores for the 26 portfolio papers individually reviewed by each of the authors; no other evaluators were involved in this review process. In this study, items that had low reliability were considered reflective of weaknesses in the RN-to-BSN curriculum, but also highlighted differences among faculty in their interpretation of the characteristics considered evidence of the attainment of specific behaviors by students.

Conceptual Framework

The conceptual basis for portfolio evaluation in nursing education is rooted in three distinct yet interrelated theories:

- Kolb’s (1984) experiential learning theory.
- Mezirow’s (2000) transformation learning theory.
- Benner’s (2001) model of the levels of proficiency in nursing.

Experiential learning and transformation learning are considered integral to the development of nursing expertise. To move from one level of proficiency to the next, a student must be willing to engage in concrete clinical practice experiences that provide professional socialization and development, as well as challenge his or her preconceptions and assumptions of competent patient care. Critical reflection through the creation of a portfolio enables the student to analyze the actions and skills used in clinical situations, cope with the emotional responses that arise from specific experiences, feel a sense of achievement, appreciate the level of responsibility inherent in the practice of nursing, and recognize the limits of formal knowledge (Benner, Tanner, & Chesla, 1996). This process also aids the student in discovering how he or she best learns, and...
what is important to learn, to achieve expertise within the context of one’s own professional practice (Daley, 1999). Thus, the goal of portfolio evaluation is to discover, document, and rate the evidence of students’ clinical knowledge and skills that have been transformed through experience and reflection.

**METHOD**

Development and testing of the PET occurred in two phases. Phase one involved development of the PET, whereas phase two included faculty training on the portfolio review process followed by testing of the PET with the evaluation of student portfolios.

**Phase One: Development of PET**

The author served as the portfolio coordinator and chaired the faculty committee charged with implementing portfolio evaluation in the baccalaureate nursing program. A seven-step process was used to develop the PET.

*Review of Examples of Scoring Rubrics.* Examples of scoring rubrics used in higher education, particularly nursing, as well as in elementary and secondary education, were obtained and reviewed by the faculty committee charged with developing the PET. The diverse formats enabled a faculty committee to make inferences regarding how these sample rubrics might be used to guide the creation of a PET for summative evaluation in a nursing program. Specific characteristics of a good PET, meaning an instrument that would facilitate consistent observations by faculty, were identified and included:

- Space for written comments.
- A rating scale with clear definitions for each level.
- Concrete examples of defining characteristics for each evaluation criterion.
- Clear differentiation between criteria related to professional competencies and criteria related to the quality or technical aspects of the portfolio.
- Provisions for a penalty for late submission of the portfolio and lack of evidence of required components.

*Review of Program Outcome Statements for Suitability as Portfolio Evaluation Criteria.* Program outcome statements were intended to be used as portfolio evaluation criteria. However, consensus of the faculty committee was that some of the outcome statements were too vague and did not relate to the definitions for curricular concepts approved by the total school faculty; thus, these statements were difficult to measure through portfolio evaluation. In this school of nursing, course objectives and program outcomes relate to seven identified curricular concepts including critical thinking, therapeutic nursing interventions, communication, teaching, research, leadership and professionalism, and standards of practice. Continued discussion on the desired qualities that should be exhibited in the student portfolio, as evidence of the competencies of the novice professional nurse, guided revision of the end-of-program outcome statements that would also serve as evaluation criteria for section 1 of the PET.

*Identification of Performance Criteria to Be Evaluated.* A set of defining characteristics for each end-of-program outcome statement was compiled from performance criteria listed in *The Essentials of Baccalaureate Education for Professional Nursing Practice* (American Association of Colleges of Nursing, 1998) The challenge for the committee was to identify performance criteria that would correlate with each of the school’s end-of-program outcome statements to create a list of 3 to 5 defining characteristics for each program outcome.

*Identification of Technical Aspects of the Portfolio to Be Evaluated.* Technical criteria that faculty believed should be incorporated into the PET were discerned from the literature but also included factors that were particularly important to faculty (e.g., the quality of reflection). When standards for technical criteria could not be determined from the literature, the faculty group created its own list of 3 to 7 defining characteristics for each of these criteria that would make up section 2 of the PET, including:

- Use of self-reflection and self-evaluation.
- Compliance with required components of the portfolio.
- Participation in educational, professional, or community service activities.
- Development of introductory and cover pages for portfolio entries.
- Overall organization of the portfolio.
- Creativity in the development of the portfolio.
- Professionalism.
- Adherence to portfolio due date.

*Review of the Required and Optional Components of the Undergraduate Nursing Student Portfolio.* Each item on a previously developed checklist (Karlowicz, 2000) was reviewed to assure that the activity or assignment would provide the needed documentation to judge the attainment of the outcome criteria, based on the defining characteristics identified on the PET. This review helped the faculty committee articulate the desired qualities that should be exhibited in the portfolio; it also enabled the group to identify additional assignments that could be selected by students to fulfill specific requirements for documentation.

*Development of a Draft PET.* A draft PET was developed and consisted of two parts: the portfolio scoring guide and the list of portfolio evaluation criteria. The main feature of the scoring guide was the rubric that provided a numeric rating and corresponding definitions for each scoring level. Descriptors for the quality of the evidence in the portfolio were based on a 5-point Likert-type scale and ranged from 1 (no evidence of the criterion) to 5 (strong evidence with concrete examples of the criterion). Descriptors of the extent to which the defining characteristics for each criterion are represented in the documentation were defined as a percentage of the total number of defining characteristics represented by entries in the portfolio and ranged from less than 20% (a rating of 1) to 100% (a rating of 5) (Table 1). Specific portfolio evaluation criteria were divided into two groups with end-of-program criteria (a total of seven items) comprising section 1 of the PET; and technical aspects of the portfolio (a total of six items) comprising section 2 of the
PET (Table 2). Within these sections, each criterion, along with its list of defining characteristics, was printed separately; space for written comments specific to each criterion was also provided. The last page of the evaluation tool was designed to provide students with a summary of the ratings for each section and calculation of an overall portfolio score. Space for final comments about the portfolio by the evaluator was also provided.

Establishment of Face Validity. Internal and external reviews of the draft PET were conducted to validate its content and design. Internal review of the PET was conducted by a committee consisting of five undergraduate faculty members. These reviewers expressed confidence that the definitions contained in the scoring rubric would enable them to assign an objective rating to evaluation criteria. Directions for rating portfolio criteria were considered succinct and offered an easily understood overview of the evaluation process. Within the main body of the evaluation tool, there were recommended changes in the wording of some of the defining characteristics. Additional behaviors valued by the faculty were fashioned into defining characteristic statements and added to lists for the appropriate outcome criteria. The changes enabled the faculty committee to affirm that the lists of defining characteristics were statements representative of the criteria being evaluated.

After making revisions to the PET based on feedback from internal reviewers, an external review of the PET was conducted by three individuals with expertise in portfolio evaluation. One committee member was external to the nursing department but a member of the university community, whereas the other committee members held positions at universities outside the state. The reviewers praised efforts to limit the subjectivity in the assignment of a score by providing a set of defining characteristics by which to judge each criterion.

**Phase Two: Faculty Training and Testing**

**Faculty Training.** To serve as an evaluator, faculty members were required to participate in departmental training conducted by the portfolio coordinator. Training was conducted in two separate sessions. The first session acquainted faculty with the portfolio development process, including the checklist of required and optional components of the portfolio. Portfolios completed by previous graduates of the undergraduate nursing program were shown to faculty to familiarize them with the various presentation and organization styles of student portfolios. During the first session, the PET was also presented and reviewed. Background on the development of the tool was provided along with an explanation of how to complete a systematic review of a portfolio and apply the scoring rubric to rate evaluation criteria. Directions provided for rating portfolio criteria included the following steps:

- Look at the presentation and overall organization of the portfolio.
- Verify that the portfolio contains all of the required components.
- Conduct a review of all required and student-selected entries in the portfolio.
- Use the rating scale to assign a rating to each criterion based on the overall quality of the evidence and the extent to which the defining characteristics for each criterion are represented in the documentation.
- Add the ratings from section 1 (end-of-program outcomes) and section 2 (technical criteria); the sum of both sections determines the overall portfolio score.

Faculty evaluators were reminded that portfolio evaluation is intended as a summative evaluation process to determine whether the baccalaureate student has progressed toward the role of professional nurse upon graduation. Evaluators were not to regrade projects and assignments included in the portfolio. Instead, they were asked to judge whether specific competencies had been attained through the validation of evidence contained within the portfolio.

During the time between the first and second training sessions, faculty members were asked to conduct a practice review of two portfolios; all faculty members reviewed the same two portfolios. Completed PETs were discussed at

---

**TABLE 1**

Portfolio Evaluation Tool Scoring Guide

<table>
<thead>
<tr>
<th>Rating Scale</th>
<th>Quality of Evidence</th>
<th>Percentage of Defining Characteristics Represented</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Documentation provides no evidence of criterion</td>
<td>&lt; 20%</td>
</tr>
<tr>
<td>2</td>
<td>Documentation provides questionable evidence of criterion with vague examples of learning experiences</td>
<td>40%</td>
</tr>
<tr>
<td>3</td>
<td>Documentation provides inconsistent evidence of criterion with some concrete and some vague examples of learning experiences</td>
<td>60%</td>
</tr>
<tr>
<td>4</td>
<td>Documentation provides consistent evidence of criterion with concrete examples of learning experiences</td>
<td>80%</td>
</tr>
<tr>
<td>5</td>
<td>Documentation provides strong evidence of criterion with concrete examples of learning experiences</td>
<td>100%</td>
</tr>
</tbody>
</table>
the second training session, which was held 3 to 4 weeks after the first session. At that time, specific issues and questions related to the review process were addressed; these included offering clarification on how to judge the quality of the evidence in the portfolio and how to apply the scoring rubric. Given that the PET has space for evaluator comments specific to each criterion, time was also devoted to observations that deserve a written comment.

**Portfolio Review Process.** At the time of this study, six faculty members had completed training to evaluate portfolios. Each student portfolio was assigned to 2 of the 6 faculty evaluators for review; portfolio review assignments were random to reduce rater bias. During a 3-week period, evaluators independently conducted reviews of the portfolios assigned to them. Completed PETs were submitted to the portfolio coordinator who reviewed all documents to ensure they had been filled out completely and correctly.

**Sample.** Evaluation of the effectiveness of the PET was conducted on a convenience sample of portfolios created and submitted by a total of 60 graduating baccalaureate nursing students in one school of nursing. Thirty-four students were in the May graduation group, and 26 students were in the August graduation group the year this study was conducted.

**Statistical Analysis.** Following approval of the project by the university’s Institutional Review Board, a database of portfolio scores was created to facilitate analysis using the SPSS version 10.0.5 statistical software program. A numeric identifier was assigned for each portfolio reviewed and scored (hereafter referred to as a *case*); similarly, a numeric identifier was assigned to distinguish between evaluators. Scores generated by the “A” and “B” rater for each case were entered into the database. There were a total of 16 individual scores per case, which included the scores assigned to the seven portfolio evaluation criteria in section 1 of the PET pertaining to the end-of-program outcomes, and the overall or final portfolio evaluation scores by the pairs of faculty evaluators in 60 cases. An evaluation of scores for technical criteria in section 2 of the PET was not included in this study. Analysis included the following:

- Determination of the point difference distribution in the final portfolio scores rendered by each pair of faculty evaluators in 60 cases (i.e., final score by rater A minus final score by rater B).

---

**TABLE 2**

**Examples of Portfolio Evaluation Criteria and Defining Characteristics**

<table>
<thead>
<tr>
<th>Criteria No. 4, Teaching</th>
<th>Rating&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Outcome</strong></td>
<td></td>
</tr>
<tr>
<td>Uses teaching strategies to maximize client health and enhance professional development</td>
<td></td>
</tr>
<tr>
<td><strong>Defining characteristics</strong></td>
<td></td>
</tr>
<tr>
<td>Provides teaching to patients or professionals about health care procedures and technologies in preparation for and following nursing or medical interventions</td>
<td></td>
</tr>
<tr>
<td>Provides relevant and sensitive health education information and counseling to patients and families, in a variety of situations and settings</td>
<td></td>
</tr>
<tr>
<td>Uses information technologies and other appropriate methods to communicate health promotion, risk reduction, and disease prevention across the life span</td>
<td></td>
</tr>
<tr>
<td>Evaluates the efficacy of health promotion and education modalities for use in a variety of settings with diverse populations</td>
<td></td>
</tr>
<tr>
<td>Uses information technologies and other appropriate methods to enhance one’s own knowledge base</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Criteria No. 12, Reflection</th>
<th>Rating&lt;sup&gt;b&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Outcome</strong></td>
<td></td>
</tr>
<tr>
<td>Uses a process of self-examination to facilitate personal and professional growth and development over time</td>
<td></td>
</tr>
<tr>
<td><strong>Defining characteristics</strong></td>
<td></td>
</tr>
<tr>
<td>Engages in continuous and complex self-reflection and self-evaluation of performance</td>
<td></td>
</tr>
<tr>
<td>Recognizes personal strengths and abilities</td>
<td></td>
</tr>
<tr>
<td>Summarizes accomplishments</td>
<td></td>
</tr>
<tr>
<td>Acknowledges areas of weakness</td>
<td></td>
</tr>
<tr>
<td>Prescribes future learning needs</td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup> This is an example from Section 1 of the Portfolio Evaluation Tool.
<sup>b</sup> Rated on a scale from 1 to 5.
<sup>c</sup> This is an example from Section 2 of Portfolio Evaluation Tool.
Determination of interrater reliability coefficients among final scores in 60 cases.

Comparison of the group mean of final scores rendered by all of the A raters versus all of the B raters for the 60 cases using the paired-samples t test.

Determination of reliability among all ratings for the seven end-of-program outcome criteria in 60 cases and for the subgroup of portfolios evaluated by each pair of faculty evaluators, by computing interrater reliability coefficients.

Determination of general tendencies in score assignment by calculating individual faculty evaluators’ mean scores for each of the seven end-of-program outcome criteria from all portfolios evaluated.

Written comments in the completed PET were also reviewed to identify remarks that might provide insight on why a faculty evaluator assigned a particular score for a criterion. Although all 60 pairs of PETs were reviewed, particular attention was paid to PETs in which there was a significant difference between the scores assigned to a criterion by both faculty evaluators analyzing the portfolio.

RESULTS

Analysis of Ratings

The maximum score on the PET was 50 points, with scores for all cases ranging from a minimum of 29.2 to 50.0 points (mean = 45.03±4.05). The mean final score for all portfolios evaluated in May was 44.80±4.31, whereas the mean final score for all portfolios evaluated in August was 45.37±3.70.

When the point differences in final portfolio scores for all 60 cases were analyzed, results showed that the difference was less than 1 point in 19 of the 60 cases and between 1.1 and 2.0 points in 12 of the 60 cases. Thus, in 52% of all portfolios evaluated, the difference in final scores between the two raters was less than 2 points. By comparison, 27% of cases had a final score point difference of 2.1 to 4.0 points, and 22% of cases had a final score point difference greater than 4.1 points.

Analysis of interrater reliability in the 60 pairs of final portfolio evaluation scores revealed a reliability coefficient of 0.78. Further analysis, using a paired-samples t test, compared the group mean of final scores assigned by all of the A raters (44.6±4.2) versus all of the B raters (45.5±3.9) for the 60 cases. Results revealed there was no significant difference in the paired scores (p = 0.07) (Table 3).

Analysis of the ratings for the seven end-of-program outcome criteria among the 60 cases revealed a reliability coefficient of 0.81. When the scores for the seven end-of-program outcome criteria were analyzed for each subgroup of portfolios reviewed by the different pairs of faculty evaluators, reliability coefficients were found to range from a low of 0.46 to a high of 0.98 (Table 4).

In an effort to determine general tendencies in the assignment of scores, raters’ mean scores for each of the seven end-of-program outcome criteria were calculated. Graphic analysis of the score means for each evaluator showed that the ratings by all faculty evaluators tended to cluster between 4.0 and 5.0 for all criteria, except for criteria 7. Analyses of raters’ mean scores for criteria 7, which addressed standards of practice, revealed that the mean scores for 5 of the 6 evaluators clustered between 4.0 and 4.5; the sixth evaluator tended to rate this criteria much lower than the other raters, with a mean score of 3.08 (Figure).
istics not represented in the portfolio documentation. Of note, evaluators consistently used the comment section to document observations when the score assigned was less than the highest rating possible for a criterion.

In one instance, notations made by a rater pair revealed conflicting impressions of the quality of the evidence contained in the portfolio for the outcomes related to communication and research. There was also disagreement about the defining characteristics not represented in the documentation. Despite these differences, the written comments regarding the quality of the evidence supported the scores assigned by each rater.

In another instance, two different pairs of raters had significant variation in the scores assigned for the outcome criteria related to standards of practice. In both rater pairs, the A rater (each pair had the same A rater) seemed to apply the scoring guide correctly based on the written comments provided. On the other hand, the B rater in each of these pairs (each pair had a different B rater) seemed to assign a score that would account for some deficiency in the portfolio evidence, but the score assigned did not match the comments and did not reflect correct application of the scoring rubric. In other words, the B raters assigned a “sympathy score” that reflected a deficiency but was not as low as the score that should have been assigned had the scoring rubric been correctly applied.

### Discussion

An effective PET should guide its users to appropriately assign ratings based on objective observations each time a portfolio is evaluated. Although interrater agreement is considered the prime indicator of the quality of portfolio evidence, as well as an indicator of the consistency in analysis among raters, there is some question about whether raters’ scores need to be the same or nearly the same. This study demonstrated that rater agreement approximating 100% is possible, although not easily achieved and perhaps not necessary as long as scores assigned for criteria by each portfolio rater are within 1 point. Attaining consensus among raters depends on a number of factors, including uniformity in the number and kind of portfolio entries, a refined scoring rubric with clear definitions of performance that everyone understands, and the honed skills of thoroughly trained evaluators that are developed through continual involvement in portfolio review and analysis.

Reliability coefficients above 0.80 are considered strong evidence of agreement among portfolio raters. The strong reliability coefficients achieved in this study are attributed to the quality of entries in the student portfolios; these are based on portfolio development guidelines that identify required and recommended items that have the potential to demonstrate the skills or competencies being evaluated. The apparent consistency with which raters applied the scoring rubric also contributed to the reliability in this study. Although scores among the pairs of raters were not identical, there was only a slight variation in raters’ mean scores, thus demonstrating the effectiveness of the PET to guide faculty in making similar judgments about the evidence offered by students in support of the attainment of program outcomes.
The results of this study revealed possible sources of unreliability in portfolio scoring. However, the recommendations for improving reliability do not represent an overhaul of the process, but rather refinements in the PET and increased faculty training with emphasis on correct application of the scoring rubric. In particular, the inconsistent number of defining characteristics for criteria was considered problematic and complicated the score assignment. Thus, it was determined that five defining characteristics for each evaluation criteria were needed to correlate with the scoring rubric, which uses a 5-point scale. Evaluators believed that this change alone would facilitate greater consistency in score assignment. Consequently, changes similar to this should be viewed as an opportunity to enhance the evaluators' understanding and expectations of student performance, as well as the meaning of portfolio evaluation scores.

It is important to keep in mind that instruments developed for portfolio evaluation differ from other measurement tools in that evaluation criteria cannot be standardized. Although schools may identify similar curricular concepts, the definitions created will reflect the faculty's unique interpretation of those concepts. Similarly, program outcome statements for each school differ and are based on the teaching philosophy of faculty, theoretical framework of the curriculum, institutional mandates, and accreditation guidelines, as well as the professional issues faculty deem particularly important. What can be standardized, and was tested in this study, was a process for assigning scores to portfolio evaluation criteria using a 5-point scoring rubric and an instrument format that requires five defining characteristics for each outcome statement to be used as key indicators of outcome attainment. As the results of this study demonstrated, a systematic approach to portfolio evaluation provides an effective framework that enables trained faculty raters to consistently make similar judgments about the quality of evidence in nursing student portfolios. Thus, the methods used to create the PET and the process designed to facilitate scoring is a model for portfolio evaluation that could be used by any school.

LIMITATIONS

This study was limited because it was valid only for the PET developed for one school of nursing. This study was also limited because portfolios were reviewed by a pool of six trained faculty evaluators, with each different pair of faculty responsible for scoring two to eight portfolios. Evaluation assignments were made to ensure that all portfolios would be reviewed by two different raters, but the number of portfolios each rater evaluated was not standardized. Finally, the study was limited because of a 3-month interval between the reviews of portfolios by May and August graduates, during which time faculty evaluators met informally to discuss issues pertaining to the review process, including the assignment of scores to evaluation criteria.

CONCLUSION

The development of an evaluation tool that will help faculty evaluators assign valid and reliable scores is a necessary first step in the effort to give meaning to the use of portfolios as a program evaluation method. Although this study has described the process for creating and testing the effectiveness of a PET to guide faculty in making consistent observations about the quality of students' work, it will be important for other programs using portfolio evaluation to test this scoring method. It will also be important to repeat the study when changes and updates to this program's curricular concepts occur to determine whether score reliability can be maintained when the evaluation criteria are changed. Although this study did not address the usefulness of portfolios for program evaluation, future research must focus on relating portfolio evaluation scores to other measures of program evaluation, as well as examining how faculty use the data generated for program improvement.

REFERENCES


