Future Academy®’s Multidisciplinary Conference

Development of an Evaluation Capacity-building Program for Nurse Teachers in Students’ Practicum

Netrachanee Kamonratananun⁎, Siridej Sujiva, Kamonwan Tangdhanakanonda

Department of Educational Research and Psychology, Faculty of Education, Chulalongkorn University, Bangkok 10330, Thailand

Abstract
Crucial problems commonly found in practicum evaluation of nurse teachers are rater effects. One among effective solutions is rater training. Despite these problems, studies in this population are scanty. The objectives of this study were to develop an evaluation capacity building program for nurse teachers in practicum evaluation of student nurses and to investigate the quality of instruments in the program. The multidisciplinary model of evaluation-capacity building was employed to develop the program. All of the faculty of nursing, Suan Dusit University and other 26 nurse teachers were recruited as participants. The results showed that instruments for the program were developed and investigated the validity, and reliability as follows 1) Rubric, IOC of 26 sub-items = .56 - .89; with corresponding 2) 8 sets of video clips, inter-rater agreement index: RAI of scores by mean RAI = .84-.95; and inter-rater reliability, ICC(3,5) 95% CI = .12 - .80 3) Manual for training, IOC = .78-.89 4) An attitude scales, IOC = .67 - 1.00; Alpha = .83 5) A test on knowledge, IOC = .44 - 1.00; Difficulty: Pi = .23 - 1.00; Sensitivity Index: Si = -.35 - .44; Alpha = .18 . After being revised some marginal values, most of the instruments were considered to be valid, and reliable except the test of knowledge. The item and test analysis were shown low to negative values in some items. However, the program will be used in a pilot group, further recalculated and improved before using in the major group.

Keywords: Rater training program; nurse teacher; practicum evaluation

1. Introduction
Evaluation of student nurses’ practicum in nursing science education is important and essential for the production of quality nurses after their graduation. However, an important problem commonly found in performance assessment

⁎ Corresponding author. Tel.: +66-2218-2565 ext. 500; Fax: +66-2218-2559; ⁎ +66-2218-2565 ext. 506, Fax +66-2218-2559
E-mail address: netrachanee_kam@dusit.ac.th
of practicum is rating errors brought about by the subjectivity of rater (Nitko & Brookhart, 2007; Wood, 1982). One of the most preferable solution is rater training (Donaldson & Gray, 2012). This paper proposed the development of a training program to enhance evaluation capacity of nurse teachers in student nurses' practicum.

2. Problem Statement

Common problems in practicum evaluation of student nurse include rater bias, rater errors, or rater effects, one of the major problems is subjectivity of the clinical nurse teacher (Wood, 1982) described in several different types (Reynolds, Livingston & Willson, 2009). Firstly, halo effect is the rater's impression towards particular individual group on particular aspect that accordingly affects the determination over other irrelevant aspects (Saal, Downey, & Lahey, 1980). Secondly, leniency/stringency error is the rater's response to the difficulty in scoring and consistently rate higher or lower by using other measures other than the defined criteria of competency level. Thirdly, central tendency error is the rater rates in the middle of scale for all in group. The errors discussed above are considered systematic error which effect on the whole group of ratees can be better statistically determined than the random error with individual but not group effect (Kane, 2011). At present, the method widely recognized and used to reduce rater effects is rater training, the original training described by Bernardin (Bernardin & Buckley, 1981). The strategies which were widely accepted for their advantage in reducing rater effect including, Rater Error Training (RET), Frame- of- Reference Training (FOR), Behavioral Observation Training (BOT), and Decision-Making Training (DMT). Each training approach provides different advantages. RET helps reduce rating error such as leniency and halo effect (Smith, 1986). FOR helps improve rating accuracy (Roch, Woehr, Mishra, & Kieszczynska, 2012). BOT approach helps increase observation accuracy (Noonan & Sulsky, 2001; O'Sullivan & Roch, 2003). DMT based on Bandura's Social Cognitive Theory (1991) clarify the nature of decision making and enhances the rater to be better in self-regulation operation when rating. Rater training is a preferred method to reduce rater error (Roch, Woehr, Mishra, & Kieszczynska, 2012) and it has also been applied in the field of medical science (Holmboe, Hawkins, & Huot, 2004; Cook, Dupras, Beckman, Thomas, & Pankratz, 2008; Ju Lin et al., 2013). However, studies on rater training, in particular, of nurses are scanty. In addition, such training should be designed to suit particular context (tailor rater training) to respond specific needs of each organization (Preusche, Schmidts, & Wagner-Menghin, 2012).

3. Research Questions

What is the model of training program to be developed for ECB among nurse teachers for practicum evaluation of student nurses?
What should be the elements in the training program?
For the training program being developed, to what extent is the quality of tools used in the training program?

4. Purpose of the Study

The objectives of the study were to 1) develop an ECB program for nurse teachers in practicum evaluation of
student nurses that can evaluate knowledge, attitude, and skills; 2) investigate the quality of the instruments in the program. The context of the study is the faculty of nursing, Suan Dusit University. The program included the development of these tools. Firstly, scoring rubric for practicum evaluation using the conceptual framework drawn from the Qualifications Framework for Thailand’s Higher Education (TQF) which contains the learning outcomes or competencies intended for students to acquire after graduation. The competencies comprise 6 domains, namely (1) ethical and moral development, (2) knowledge: the ability to understand, recall and present information including knowledge of specific facts, knowledge of concepts, principles and theories and knowledge of procedures, (3) cognitive skills: the ability to apply knowledge and understanding of concepts, principles, theories and procedures when asked to do so; and analyse situations and apply conceptual understanding of principles and theories in critical thinking and creative problem solving when faced with unanticipated new situations, (4) interpersonal skills and responsibility, (5) analytical and communication skills, (6) practical skill for the professional. The learning outcomes also include good personality and precision at work as the unique attributes of Suan Dusit University. Secondly, video clips of simulation as performance tasks for practicing evaluation and to be used with scoring rubric. Thirdly, manual for ECB training. Fourthly, the attitudes scale to assess the attitude towards practicum evaluation. Fifthly, test of knowledge about practicum evaluation.

5. Research Methods

5.1 Design
The present study was employed the research and development design, adopted Preskill and Boyle’s (2008) Multidisciplinary Model of ECB that comprised various strategies, i.e., training, coaching, involvement in an evaluation process, meeting etc., and emphasizes participatory evaluation of the faculty of nursing, Suan Dusit University, as the research setting of the major study.

5.2 Setting
The faculty of nursing, Suan Dusit University was used to survey the problems in practicum evaluation and participated in the designing and developing of rubric and video clips as the performance tasks, and a nursing college that had a similar context as the major study to try out the knowledge test, and the attitude scale on practicum evaluation.

5.3 Sample
Nurse instructors of the faculty of nursing, Suan Dusit University were used to survey the problems involving practicum evaluation. Purposively selected internal experts included 7 instructors of the faculty who have expertise and experience in teaching both theories and practices for at least 10 years, examined the instruments including content validity and determine index of item objective congruence (IOC) of rubric, video clips’ script. Thirty-five nurse instructors who had experience in practicum evaluation were recruited from a college of nursing that had similar context as the sample of the major study to be tried out for the knowledge test and attitude scales in practicum evaluation.
5.4 Measurement
Content validity was determined by Index of item objective congruence (IOC), items analysis were examined by using item difficulty ($P_i$), sensitivity index ($S_i$), whereas Cronbach's alpha was used to examine reliability. Inter-rater agreement was examined by rater agreement index (RAI). Inter-rater reliability was investigated by intra-class correlation coefficient: ICC (3,5). As stated by Shrout and Fleiss (1979) 3 stood for model 3 which each subject (video clip) was assessed by each rater who was the only rater of interest. Meanwhile, 5 represented the mean rating of 5 raters. SPSS statistical packages was used to produce p-value, $S_i$, and Alpha and Excel was used to indicated IOC and RAI.

5.5 Procedure and analysis
Different types of tool were developed as follows 1) Rubric: The researcher started with reviewing related literature regarding rubric development, using Popham's approach(1997), then surveyed the faculty of nursing on need for training and elements in the program after that created evaluating criteria and chose 4 levels of assessment as it is convenient to give the description for each level (Perie, 2008). Subsequently, the research defined each level of performance and its corresponding descriptor. Purposively selected internal experts included 7 instructors of the faculty of nursing who have expertise and experience in teaching both theories and practices for at least 10 years, to examine the content validity and determine IOC. The result was brought to revision. External experts were selected from those with educational measurement and evaluation qualification with 6 in 9 of them have knowledge background in nursing, to examine the content validity and determine IOC. 2) Eight sets of video clips as performance tasks for practicing practicum evaluation corresponding to the rubric were divided by the levels of performance. The simulations in video scripts were identified and reviewed by internal experts of the faculty of nursing, Suan Dusit University. Then the research examined the video clips and rubric congruence by 5 external experts with at least 10 years of experience in nursing education both theories and practices. The meeting date was set to find conclusions and determine the performance level for each set of video clips, and examine inter-rater agreement scores among the 5 experts (raters) on the 8 sets of video (ratees) using rater agreement index, and examine inter-rater reliability using intra-class correlation coefficient; ICC (3,5). Rubric was intended to be used for practicing evaluation skill in combination with practical tasks or the sets of video clips. 3) ECB training manual for enhancing expected basic knowledge about practicum evaluation was consisted of 6 units, e.i. introduction, basic knowledge about performance assessment, rater error training, frame of reference training, behavioral observation training and decision making training. Content validity was examined with IOC by the same experts for rubric. 4) The attitude scales, 5 point likert-type scale where 1 to 5 represented strongly disagree to strongly agree, was used to assess the attitude towards practicum evaluation at pre- and post-training. A pilot study was conducted to try out the attitude test and the test of knowledge with 35 participants recruited from the nurse instructors who had experience in practical evaluation from a college of nursing that had similar context as the sample of the major study. The returned questionnaires were 26. Content validity and reliability were respectively examined with IOC.
and Cronbach’s alpha 5) A 25 item-test of knowledge about practicum evaluation examined with IOC for content validity by the same experts for rubric, item difficulty (P_i), sensitivity index (S_i) were examined for items analysis and Cronbach’s alpha for reliability. The test, the attitude scale and the performance task corresponding with rubric were aimed to examine the knowledge, attitude, and skill at pre- and post-training for next phase of the study.

6. Findings

6.1 Results

The instruments for the programme were developed as follows: 1) Rubric for practicum evaluation containing 7 key competencies and 26 sub-items 2) Eight sets of video with performance level determined by experts 3) ECB training manual for enhancing expected basic knowledge about practicum evaluation 4) 15 attitude scales to assess the attitude towards practicum evaluation 5) A 25-item questionnaire on knowledge about practicum evaluation. The validity and reliability of instruments were investigated as follows, 1) 26 sub-items rubric for practicum evaluation with IOC of each sub-item ranged between .56 and .89, with corresponding eight sets of video, inter-rater agreement of scores of RAI for video set 1 = .68 - 1.00, set 2 = .80-1.00, set 3 = .73 - 1.00, set 4 = .73 -.86, set 5 = .80 - 1.00, set 6 = .73 -.86, set 7 = .80 - 1.00, set 8 = .80 - 1.00 shown in Mean RAI, and inter-rater reliability with ICC(3,5) 95% CI = .12 - .80 as seen in Table 1.

<table>
<thead>
<tr>
<th>Video</th>
<th>Mean RAI</th>
<th>ICC</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1</td>
<td>.84</td>
<td>.12</td>
<td>-.56 -.55</td>
</tr>
<tr>
<td>Set 2</td>
<td>.90</td>
<td>.49</td>
<td>.10 -.75</td>
</tr>
<tr>
<td>Set 3</td>
<td>.92</td>
<td>.80</td>
<td>.61 -.90</td>
</tr>
<tr>
<td>Set 4</td>
<td>.85</td>
<td>.63</td>
<td>.35 -.82</td>
</tr>
<tr>
<td>Set 5</td>
<td>.95</td>
<td>.55</td>
<td>.13 -.78</td>
</tr>
<tr>
<td>Set 6</td>
<td>.84</td>
<td>.45</td>
<td>-.01 -.71</td>
</tr>
<tr>
<td>Set 7</td>
<td>.88</td>
<td>.32</td>
<td>-.30 -.67</td>
</tr>
<tr>
<td>Set 8</td>
<td>.94</td>
<td>.36</td>
<td>-.23 -.69</td>
</tr>
</tbody>
</table>

n = 5

3) Manual for ECB training with IOC = .78-.89. The 26 participants, the test takers for the attitude scales and knowledge test, 100% of them had a background in nursing, 92.30% were female, 100% held a master degree or above, and none of them has never trained to decrease rater effects in practicum evaluation. The mean age was 37.62 (range 31-52 years) 4) A 15 items to assess the attitude towards practicum evaluation, with IOC ranged between .67 and 1.00, Cronbach’s alpha = .83 as presented in Table 2.
Table 2

<table>
<thead>
<tr>
<th>Statement</th>
<th>IOC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. I feel good in preparing for teaching practicum.</td>
<td>1.00</td>
</tr>
<tr>
<td>2. I think it is very easy to prepare tasks for using in practicum evaluation.</td>
<td>.67</td>
</tr>
<tr>
<td>3. I think the use of rubric for practicum evaluation is necessary.</td>
<td>.78</td>
</tr>
<tr>
<td>4. I feel good to involve in practicum evaluation.</td>
<td>1.00</td>
</tr>
<tr>
<td>5. I feel that practicum evaluation is boring.</td>
<td>1.00</td>
</tr>
<tr>
<td>6. If possible, I want other person instead of me to involve in practicum evaluation.</td>
<td>.89</td>
</tr>
<tr>
<td>7. I think practicum evaluation wastes both time and manpower resources.</td>
<td>1.00</td>
</tr>
<tr>
<td>8. Practicum evaluation of nursing students has complicated process.</td>
<td>.89</td>
</tr>
<tr>
<td>9. In practicum evaluation of nursing students, it is very difficult to bring common understanding and scores agreement between teacher raters.</td>
<td>1.00</td>
</tr>
<tr>
<td>10. Solving problems with practicum evaluation is idealistic.</td>
<td>.78</td>
</tr>
<tr>
<td>11. It is necessary to enhance evaluation capacity of nurse teachers in practicum evaluation of nursing students through training.</td>
<td>1.00</td>
</tr>
<tr>
<td>12. Knowledge and skill practice on practicum evaluation will enhance my knowledge for better practicum evaluation.</td>
<td>1.00</td>
</tr>
<tr>
<td>13. Skill practice in practicum evaluation through simulation videos will enhance my skill for better practicum evaluation.</td>
<td>1.00</td>
</tr>
<tr>
<td>14. Training will increase my confidence in practicum evaluation.</td>
<td>1.00</td>
</tr>
<tr>
<td>15. Training will increase my practicum evaluation skill.</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Cronbach’s alpha = .83

5) A 25-item questionnaire on knowledge about practicum evaluation with IOC ranged between .44 and 1.00, with range of Item Difficulty (Pi) = .23 - 1.00, with Sensitivity Index: Si ranged between -.35 and .44 as shown in Table 3

Table 3

<table>
<thead>
<tr>
<th>Statements</th>
<th>IOC</th>
<th>Pi</th>
<th>Si</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. The key principle in practicum evaluation is to use personal experience in practical session of instructor to judge the practical learning outcome of learners.</td>
<td>1</td>
<td>.92</td>
<td>.01</td>
</tr>
<tr>
<td>2. The important steps to design the performance assessment are to set the purpose of the assessment, to identify the framework and to specify the test specification.</td>
<td>1</td>
<td>1.00</td>
<td>.00</td>
</tr>
<tr>
<td>3. Nursing students’ practicum in patient wards is considered a naturally occurring performance task.</td>
<td>.78</td>
<td>.38</td>
<td>.03</td>
</tr>
<tr>
<td>4. In case of high risk performance should not use the paper - pencil test to evaluate the knowledge before practicum.</td>
<td>.67</td>
<td>.42</td>
<td>.07</td>
</tr>
<tr>
<td>5. Checklist comprises lists of behaviors presented step by step for a rater to tick in the blank boxes whether the 6eha has or does not have those behaviors without descriptions for each performance level.</td>
<td>1</td>
<td>.73</td>
<td>-.02</td>
</tr>
<tr>
<td>6. Rating scale differs from check lists in that it can identify performance levels but not behaviors.</td>
<td>.89</td>
<td>.65</td>
<td>.28</td>
</tr>
<tr>
<td>7. Paper and pencil test for knowledge is one of instruments for performance assessment.</td>
<td>1</td>
<td>.58</td>
<td>.01</td>
</tr>
<tr>
<td>8. Examining inter-raters reliability is to examine inter-raters agreement of results.</td>
<td>1</td>
<td>.81</td>
<td>-.41</td>
</tr>
<tr>
<td>9. Creating common understanding among raters in the use of practicum evaluation tools such as scoring rubric is one solution to practicum evaluation problem.</td>
<td>1</td>
<td>.38</td>
<td>-.04</td>
</tr>
<tr>
<td>10. Rater error training includes only raters with serious errors to be trained for better quality of evaluation.</td>
<td>.89</td>
<td>.96</td>
<td>.30</td>
</tr>
</tbody>
</table>
11. Stringency error refers to rater's bias over particular ratee and rating him or her lower than actual performance. .89 .31 .05
12. Training on reference framework is to create common understanding among raters in the use of evaluation tools such as rubric or standards used as a criteria for the same evaluation in the same practicum evaluation. 1 .38 .17
13. Using statistical method is the best way to solve the rater effects. 1 .88 -.01
14. Behavioral observation practice is to improve skills in 1) searching; 2) perceiving; and 3) recalling memory on performance behavior. 1 .88 -.35
15. Training on decision making for practicum evaluation has a disadvantage in that it takes longer time for rating. .44 .54 -.23
16. Which statement is an example of task assigned by teachers to students in practical session. .67 .50 -.31
17. According to the graph of practicum evaluation results, what type of rater error displayed by the second teacher? .89 .42 .44
18. Which statement presents the major cause of error in practicum evaluation? 1.00 .62 .10
19. What type of tool is accepted for its most improving inter-raters (teachers) congruence or agreement of scores on the same student) objectivity (in practicum evaluation)? .89 .69 .05
20. Which statement is the examination of quality of practicum evaluation? .89 .23 .22
21. What type of grading is it by converting mean and standard deviation of scores of all students into standard scores to identify student's position? 1.00 .58 .07
22. Rater error or rater effect refers to error occurred by what cause? .89 .73 .13
23. What type of solution will help solving rater error most directly to the point? .78 .69 .20
24. Which statement is a good principle of behavioral observation? .89 .38 .28
25. Which statement is the situation in which practical session teacher use appropriate rules for determination? .78 .81 .06

n = 26                                                                                                                                      Mean = 15.5 , SD = 2.44     Cronbach's alpha=.18

6.2 Discussion
The development of training program was completed, the validity and reliability of 5 instruments including the followings: 1) Rubric for practicum evaluation containing 7 key competencies and 26 sub-items provided scores with IOC of each sub-item ranged between .56 and .89 that meant all of them were congruent with the objectives, in keeping with Hemphill and Westie (1950) values of the index .50 or more would be at a minimum of the perfect rating among specialists. Combining rubric with 8 video clips offered mean RAI = .84 - .95, according to Burrystock, Shaw, Laurie and Chisom (1996), the index closer to 1.00 indicates more consistent among the raters or has high level of inter-rater agreement. As reported by Cicchetti (1994), the inter-rater reliability with ICC = .12 - .80 interpreted that ICC values counted for poor inter-rater reliability (IRR) for video set 1 and 8, fair for video set 2,5 ,and 6, good for video set 4 , and excellent for video set 3. The video sets with ICC values from good to poor IRR were specified for improvement and revised. 3) Manual for ECB training with IOC = .78-.89 indicated that the manual was congruent with the objectives. 4) A 15-item questionnaire to assess the attitude towards practicum evaluation, with IOC ranged between .67 and 1.00 were interpreted as well congruent with the objectives, Cronbach's alpha = .83, referred to highly inter-item correlation as stated by Salvucci, Walter, Conley, Fink, and Saba (1997). 5) A 25-item questionnaire on knowledge about practicum evaluation, with IOC ranged between .44 and 1.00 , the value lower than .50 were revised to conform to the experts, with range of item P_i = .23-1.00, based on a rule of thumb as mentioned by Bachman (2004) p-values between .20 and .80 were selected for test takers, those
with higher than .80 or lower than .20 were revised, even if these values can be vary by group. Sensitivity Index: $S_i$ ranged between -.35 and .44, according to the criteria of the test’s discrimination, the items at lower to negative values could not be able to discriminate test taker in high, moderate, or low group (Oller, 1979). The values higher than .20 were acceptable, those with lower than .20 were revised. The test reliability: Cronbach’s alpha = .18, in accordance with Salvucci, Walter, Conley, Fink, and Saba (1997), the reliability value less than .50 is considered low, interpreted that the items poorly correlated to each other.

Most of discrimination values had low to no discrimination including low Alpha, contrasted to the IOC values which revealed perfectly congruent with the objectives among experts. Negative value of discrimination commonly occurs with ambiguity or complexity of material written and test taker in the lower group get more an item correct than the upper group (Matlock-Hetzel, 1997). Inspection the negative values, for example, in item 5 or 8, the statements were obviously written and shown the main idea of the instrument used in performance assessment in item 5 or inter-rater reliability’s definition in item 8. In addition, the background of the test takers in trial group revealed that none of them has ever learned about rater effects in practical evaluation. It may be that the basic knowledge of performance assessment or practical evaluation was a new thing for them. Using Thai library integrated system, the online university-network database system in Thailand providing educational information, found out for some key word i.e., “rater training” the result was that scanty study could be found. Therefore, all of them were assumed to be novice. On the other hand, it was the misfit of items difficulty with the ability of the testees that lead low reliability (Henning, 1987).

The reliability of the knowledge test suggested fairly low. It may occur in situations that the test length is short, or the test items are not interrelatedness (Schmitt, 1996). For this paper, the short-length test and the interrelatedness should be responsible for the low values, and some reasons could be explained that the pre- and post-test should be develop to take between 10-25 minutes to complete. Therefore, it is unreasonable to increase the test length (I-TECH, 2010). Using the test to evaluate the basic knowledge were composed of multidimension as mentioned in the problem statement session, even though a multidimensional test has been well documented it does not necessary have a lower Alpha than that of unidimension. Therefore, Alpha should be calculated for each of the concepts rather than for the entire test or scale (Tavakol & Dennick, 2011). Supposedly, we had separated the test items in unidimension each of this could contain only 1-5 sub-items that resulting in unavoidable low reliability.

Following the guideline for pre-and-post testing of the International Training & Education Center for Health (2010), validation can take place via a few testees to take the test, ask them after the post-test for feedback, then improve the test.

7. Conclusions

In conclusion, the tailor rater training program was developed to fit the particular context with the instruments that can evaluate knowledge, attitude and skills, and served the faculty’s special needs. The rubric and eight sets of video clips for skills training in practical evaluation, the ECB training manual, the attitude scales to assess the attitude
towards practicum evaluation could be considered as valid and reliable instruments. The item and test analysis of the knowledge test shown low to negative values in some items, because of new things for the testees and the limited test length for typical pre- and - post test . The test will be used in the pilot group, further recalculated, and could be improved before used in the major group.

Acknowledgements

The researchers would like to thank THE 90th ANNIVERSARY OF CHULALONGKORN UNIVERSITY FUND (ratchada phisek somphot Endowment Fund) and CONFERENCE RRANT FOR Ph.D. STUDENT for funding this research.

References


