Journal of Educational Evaluation for Health Professions

J Educ Eval Health Prof 2015, 12: 1 • http://dx.doi.org/10.3352/jeehp.2015.12.1

Open Access

eISSN: 1975-5937

RESEARCH ARTICLE

Validation of a clinical critical thinking skills test in nursing

Sujin Shin[1]* , Dukyoo Jung[2] , Sungeun Kim[2]

[1]Department of Nursing, Soonchunhyang University College of Medicine, Cheonan, Korea; [2]Division of Nursing Science, Ewha Womans University College of Health Science, Seoul, Korea

Abstract

Purpose: The purpose of this study was to develop a revised version of the clinical critical thinking skills test (CCTS) and to subsequently validate its performance. **Methods:** This study is a secondary analysis of the CCTS. Data were obtained from a convenience sample of 284 college students in June 2011. Thirty items were analyzed using item response theory and test reliability was assessed. Test-retest reliability was measured using the results of 20 nursing college and graduate school students in July 2013. The content validity of the revised items was analyzed by calculating the degree of agreement between instrument developer intention in item development and the judgments of six experts. To analyze response process validity, qualitative data related to the response processes of nine nursing college students obtained through cognitive interviews were analyzed. **Results:** Out of initial 30 items, 11 items were excluded after the analysis of difficulty and discrimination parameter. When the 19 items of the revised version of the CCTS were analyzed, levels of item difficulty were found to be relatively low and levels of discrimination were found to be appropriate or high. The degree of agreement between item developer intention and expert judgments equaled or exceeded 50%. **Conclusion:** From above results, evidence of the response process validity was demonstrated, indicating that subjects respondeds as intended by the test developer. The revised 19-item CCTS was found to have sufficient reliability and validity and will therefore represents a more convenient measurement of critical thinking ability.

Key Words: *Intention; Judgment; Nursing students; Reproducibility of results; Thinking*

INTRODUCTION

The need for critical thinking in the field of nursing has recently been emphasized, resulting in the proliferation of pertinent studies [1,2]. The Korea Institute of Curriculum and Evaluation defines the concept of critical thinking as thinking intended to grasp the logical structure and meaning of texts in order to make best judgments concerning concepts, criteria, contexts, and methods so as to decide whether to accept certain opinions or whether to conduct certain acts [3]. However, the current measurements used to evaluate general critical thinking skills or disposition levels do not adequately assess these skills in the context of the problems faced in clinical practice. Furthermore, critical thinking skills are dependent on the specific conditions and context of the field or time period. Existing studies are limited in that they examine general critical thinking skills using instruments that fail to account for the context of clinical conditions. Although critical thinking as a concept is a key objective within nursing education and practice, few standardized instruments have been developed to measure critical thinking levels specifically for the field of nursing. Thus, there is a need to look beyond a purely theoretical understanding of critical thinking and to examine the application of critical thinking processes in a more appropriate context. Simply put, an instrument needs to be developed that can measure critical thinking skills while accounting for specific geographical, cultural, and clinical contexts.

As a result of this need to develop a more refined instrument,

*Corresponding email: ssj1119@sch.ac.kr

For the revised version of the clinical critical thinking skills test, please contact corresponding author.

Received: August 25, 2014; Accepted: January 23, 2015;
Published: January 27, 2015

This article is available from: http://jeehp.org/

Shin et al. [4] developed a 30-item clinical critical thinking skills (CCTS) test and subsequently assessed the item difficulty, discriminant validity, internal reliability, content validity, and criterion-related validity of the instrument. However, the internal reliability was found to be a little low (Cronbach's α=0.55), possibly due to respondent fatigue as a result of the time required to respond to all 30 items (approximately 50 minutes). If true, the reliability of this tool might be enhanced through item response alternative analysis. Therefore, this study aimed to reevaluate the CCTS with the aim of creating a revised measure with fewer items and then to assess the reliability and validity of this revised instrument.

METHODS

Materials and subjects

This study is a secondary analysis of the CCTS [4]. Two hundred and eighty four nursing students participated in data collection for item analysis based on item response theory (IRT) in June 2011. The subjects of data collection related to test-retest reliability were 20 nursing college and graduate school students who sufficiently understood the purpose of the study and agreed to voluntary participation in July 2013. Nine of the subjects participated in cognitive interviews for the purpose of response process validity analysis. Study subjects for revalidation of content validity were two professors of philosophy, two professors of education, and two scholars of nursing with experience in studies related to critical thinking.

Technical information

The two-parameter normal ogive model of IRT was applied to conduct item analysis and the correlation coefficients between total scores of items were examined. Data met normality assumptions. The IRT two-parameter normal ogive model provides two item parameters (discrimination, difficulty) and tests information functions. The item parameters are used to distinguish items with poor discrimination and such items are flagged for exclusion. In addition, items with low correlations with total score may also be excluded because they likely measure different constructs. For item analysis, 28 out of 30 items were selected through content validation. Original item numbers 20 and 21 showed a low percentage of correct answers in the preliminary item analysis. Such items produce large errors in discrimination and difficulty estimation so that reliable parameters cannot be easily produced.

The content validity of the ability to reflect the areas of interpretation, analysis, inference, and evaluation, defined as the constructs of clinical critical thinking skills, in the developed items were assessed. Respondents were requested to judge and subsequently indicate the areas of clinical critical thinking ability best represented by given items. The degree of agreement between the intentions of the developers and the expert judgments were then calculated in percentages.

Cognitive interviews of students (also known as 'think-alouds' [5]) can examine how students think about, interpret, and respond to questionnaire items. Therefore, cases where the item response processes and outcomes of students coincided with item developer intention were coded as two points, cases where item response processes and outcomes partially coincided with item developer intention were coded as one point, and cases where the item response process and outcomes were not at all related to item developer intention, or the respondents answered "I don't know," were coded as zero points. The averages of coded values were calculated for each item.

Statistics

For item revision, the items were analyzed using IRT and the reliability and validity of the revised test instrument was analyzed. BILOG-MG ver. 3.0 (Scientific Software International Inc., Skokie, IL, USA) and IBM SPSS ver. 19.0 (IBM Co., Armonk, NY, USA) were used for item analysis. Internal reliability using Cronbach's α coefficient and test-retests were conducted to assess the reliability of the revised CCTS, and the correlations between scores at two time points were measured using Pearson correlation coefficients. The content validity of the revised items was calculated as a percentage of the degree of agreement between the intention of the instrument developers in item development and the judgments of six experts. The validity of the response process was analyzed using the content of qualitative data obtained through cognitive interviews on the respondents' response processes. Construct validity was tested using confirmatory factor analyses, which were conducted using the robust weighted least squares method known to be suitable for binary data [6]. Mplus ver. 6.11 (Muthen & Muthen, Los Angeles, CA, USA) and IBM SPSS ver. 19.0 (IBM Co.) were used to verify the goodness of the tests.

RESULTS

Items 1, 8, 9, and 28 showed low difficulty parameters not higher than -2.0. Twelve items showed appropriate or high levels of discrimination (discrimination parameter not lower than 0.2) [7]. The 16 items with low levels of discrimination were reviewed for deletion or revision. The discrimination parameters and item content were considered together as a group, and items 2, 16, 17, 22, 23, 24, 27, and 29 were excluded from the final version. The other two items 1 and 9 (with difficulty parameters not higher than -3.0), which were also reviewed for deletion or revision as both items showed a correct answer

percentage that exceeded 90%. As a result, item content and measured constructs were analyzed, and the relationships of these two items with other items were reviewed. Following this assessment, item 9 was excluded from the test. Although item 1 was identified as too easy, its content addressed issues regarding aging and the health of the elderly, which are highly utilizable in clinical situations. Likewise, item 6 was judged to be an important item for measuring the abilities of interpretation and analysis using contextual circumstances in clinical situations, and so these items were retained. Meanwhile, items 20 and 21, both initially included in the test instrument when it was developed in 2012, were judged to be items based on nursing knowledge and thus were excluded from the revised instrument. The results of calculations of the levels of difficulty and discrimination of the 28 items are shown in Table 1.

This instrument evaluated subjects ranging from those with low critical thinking ability to those with high critical thinking ability, and showed the maximum test information at points where subjects' ability parameters equaled -1.0. However, this instrument did not provide sufficient information for subjects with a critical thinking ability of 1.0 or higher. The test information function of the CCTS is shown in Fig. 1.

Nine items were excluded through IRT analysis. The correlations between items and total score for the 19 items included in the test instrument are shown in Table 2. Of these 19 items, 18 (item 1 excluded) showed a correlation with total score that exceeded 0.3, and all these correlations were significant at $P < 0.001$. In the case of item 1, the correlation with total score was calculated to be low compared to other items due to its high

Table 1. Levels of discrimination and difficulty according to item response theory (n = 284)

Item	Level of discrimination	Standard error	Level of difficulty	Standard error
1	0.628	0.205	-4.326	1.201
2	0.285	0.066	-0.059	0.251
3	0.312	0.069	-0.026	0.231
4	0.326	0.075	0.030	0.221
5	0.334	0.073	0.298	0.226
6	0.291	0.067	-0.391	0.262
7	0.445	0.087	-1.312	0.289
8	0.525	0.109	-2.095	0.393
9	0.483	0.118	-3.026	0.639
10	0.572	0.108	-1.413	0.269
11	0.423	0.085	-0.571	0.199
12	0.395	0.084	-0.872	0.254
13	0.359	0.077	-1.916	0.448
14	0.393	0.081	-1.079	0.285
15	0.505	0.094	-0.475	0.171
16	0.241	0.059	2.199	0.618
17	0.247	0.063	2.772	0.760
18	1.355	0.281	-1.066	0.129
19	0.423	0.085	-0.480	0.195
22	0.209	0.053	0.082	0.335
23	0.308	0.071	1.338	0.378
24	0.286	0.066	-0.866	0.316
25	0.402	0.082	-1.453	0.323
26	0.362	0.08	-0.575	0.236
27	0.250	0.065	3.022	0.827
28	0.643	0.124	-2.114	0.349
29	0.306	0.072	1.672	0.444
30	0.522	0.096	-0.879	0.197

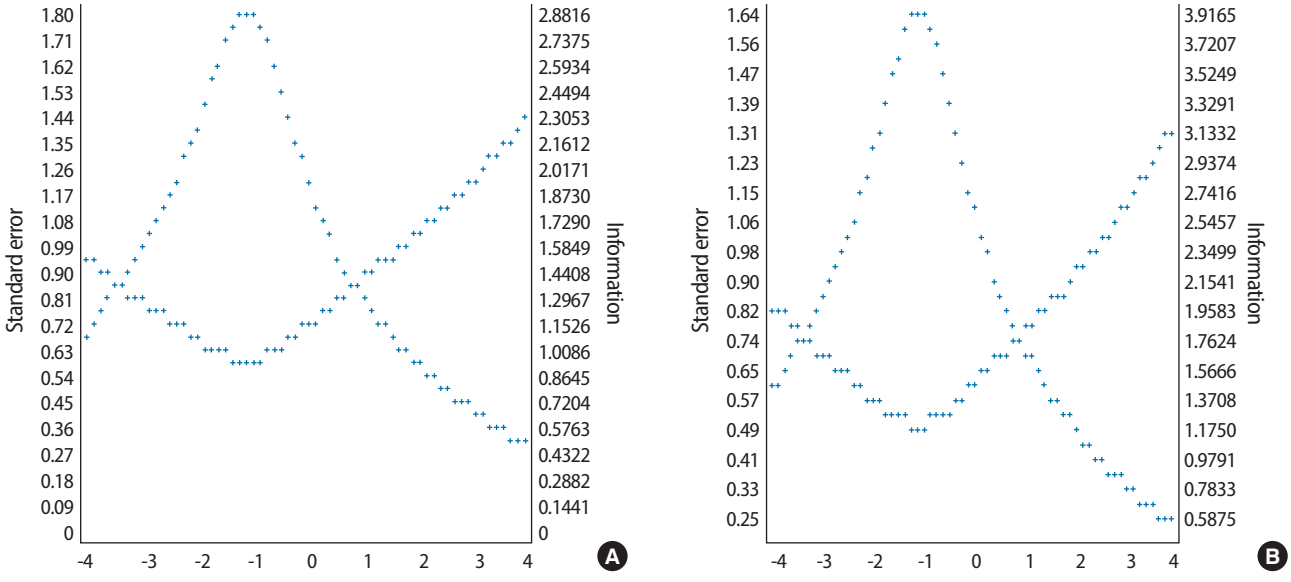

Fig. 1. Resultant test information functions. (A) Test information function of 30 items. (B) Test information function of 19 items. The test information of 30 items was peak at the -1.3 of ability parameter; while, that of 19 items was peak at the -1.1 of ability parameter. Standard error decreased in 19 items. There was higher information value in 19 items test than 30 items test.

percentage of correct answers. However, the item showed an appropriate level of discrimination and was deemed necessary to include as a result of the content analysis. Cronbach's α indicated that the reliability of the test instrument was 0.622, and the test reliability when items were removed showed a range of 0.572 to 0.623. The same 22 subjects were requested to respond to the test instrument after an interval of two weeks, and the correlations between the scores at the two time points were measured. The results showed significant correlations: r = 0.662 (P = 0 .001). The degree of agreement between item developer intention and expert judgments were calculated as a percentage for the 19 items (Table 2). Items showed agreement levels of 50% or higher. Item 7 was first developed as an analysis item, but five experts judged it an inference item and so it was eventually classified as such.

Data on the processes of thinking through which item judgments were made were collected through interviews. Most items scored at least 1.5 points, and the item scores were generally considered healthy with a total average of 1.75 points. This indicates that subjects successfully described responses as intended by the test developer. In addition, when asked the question "Was there any item you could not answer because you had no knowledge or preceding learning?" all students answered, "There was no such item." The instrument was thus verified as an instrument that measured thinking processes, not knowledge.

Confirmatory factor analyses were conducted in order to validate a model of the test instrument for measuring four factors: 'analysis,' 'understanding,' 'inference,' and 'evaluation.' Individual factors and the items for measuring the relevant factors are shown in Table 3. The goodness of fit of the confirmatory factor analyses for both the 19 items and four factors had excellent fit indices: chi-square, 77.763 (df = 69, P = 0.219); comparative fit index, 0.949; normed fit index, 0.954; and root mean square error of approximation, 0.021 since values exceeds the followings thresholds: chi-square, P > 0.05; comparative fit index and normed fit index equal to or greater than 0.9; and root mean square error of approximation equal to or less than 0.06.

Table 2. Correlations between items and total score and percentage of agreement between researcher intention and expert decision[a)]

Item no.	r	%
1	0.110	100.0
3	0.308***	100.0
4	0.309***	50.0
5	0.322***	66.7
6	0.295***	66.7
7	0.354***	16.7
8	0.354***	66.7
10	0.434***	83.3
11	0.361***	83.3
12	0.320***	83.3
13	0.338***	66.7
14	0.380***	66.7
15	0.432***	66.7
18	0.617***	100.0
19	0.356***	83.3
25	0.340***	83.3
26	0.355***	83.3
28	0.421***	100.0
30	0.409***	100.0
Total		77.1

r, correlation coefficient; %, percentage of agreement between researcher intention and expert decision.
[a)]Secondary data analysis after deleting 11 items. ***P > 0.001.

DISCUSSION

This study revised the existing 30-item CCTS instrument for clinical critical thinking ability into a 19-item measure and reported the process of instrument validation. This instrument is the first to measure critical thinking ability in the area of nursing in Korea. Unlike psychological measurements, grounds for the validity of cognitive response processes for the test instrument were set, and a new approach to expert content validity was attempted. First, the results of validation of the response processes were different from the reported levels described during interviews with the subjects. Therefore, more exploration into both difficulty and discrimination levels is considered necessary.

This study showed maximum test information at points where subjects' ability parameters were -1.0. However, the results did not provide sufficient information for subjects with critical thinking abilities exceeding 1.0, and so the instrument

Table 3. Factors and measured items

Factor name	Item no.
F1: finding the evidence and cause and evaluating	6, 8, 10, 11, 13, 14, 15, 18, 25, 28, 30
F2: interpreting and inferring the meanings	4, 5, 19
F3: inferring and evaluating the relation	3, 7, 26
F4: finding the best solution through inference and evaluation	1, 12

reported in this study is limited to use with subjects with excellent critical thinking ability scores. However, since the instrument has the advantage of identifying those critical thinking abilities necessary for medical personnel, this may be strength when used with this demographic.

Although items with positive correlation coefficients may be interpreted as measuring the same constructs as the test is intended to measure [8], this is generally considered only applicable to items with correlation coefficients exceeding 0.30. The correlations between item scores and total test score (with the exception of item 1) satisfy both criteria. This means that clinical critical thinking ability may be measured through individual items. In this study, after the number of items was reduced to 19, primarily through the selection of items with high levels of discrimination and a reorganization of the items, the reliability of the instrument was improved to 0.622. In addition, since respondent fatigue presumably decrease resulting in improved concentration following the reduction in the number of items [9], test-retest reliability showed high, statistically significant correlations ($r = 0.662$).

Whereas existing methods of verifying content validity provide information on the constructs to which items belong and evaluate the suitability of items for those constructs and content, in this study the rates of agreement between item developer intention and expert judgments were developed by having experts evaluate the content of each item for constructs. However, because of the nature of critical thinking, the subareas of interpretation (analysis, inference, and evaluation) do not act independently, but interact in order to more accurately judge given situations and to generate solutions to problems. Therefore, it is difficult to develop items that independently measure the different subareas of critical thinking skills. In this study, when the degree of agreement between the constructs to which items belonged and the constructs evaluated by experts were evaluated, most items showed agreement rates in excess of 50%. These are within acceptable parameters [10].

This research is the first among nursing studies to present evidence for response process validity. In particular, since this test is a cognitive evaluation instrument, how items are interpreted or accepted by test subjects is important [5]. Subjects' critical thinking processes were evaluated through selective items, and response processes were analyzed in order to assess whether such selective type items were well designed. Critical thinking processes are composite processes and are evaluated through multiple-choice measures and open-ended tests. Since constructed response items generally induce complicated thinking processes, while multiple-choice items typically induce low-level cognitive processes, constructed response items are able to measure cognitive processes more directly [7]. However, well-made multiple-choice measures can be useful in evaluating critical thinking ability [7,11], because judgment ability can be measured by presenting situations using item stories through selective type items and having subjects select the best response among the response alternatives presented for the specific situation [7]. Therefore, in this study, response processes were evaluated in order to determine whether the revised instrument was suitable for measuring critical thinking skills. This was determined by assessing whether subjects underwent the processes of finding responses to relevant items using the critical thinking skills intended by the developer. The high average comparison score of 1.75 supported the notion that the items were suitable for measuring critical thinking skills. These results are similar to the degree of responses for students with high levels of achievement reported in a previous study [5], in which response data regarding response processes were analyzed using a similar method. This is significant in lending further support to the validity of the revised test instrument for evaluating critical thinking ability.

Finally, confirmatory factor analyses were conducted on the validity of constructs, comparative fit index, normed fit index, and root mean square error of approximation exceeded thresholds, indicating that the collected data supported the factor model of the test. The four factors were named 'finding the evidence and cause and evaluating,' 'interpreting and inferring the meanings,' 'inferring and evaluating the relationship,' and 'finding the best solution through inference and evaluation.' These are different from the original theoretical concept subareas (interpretation, analysis, inference, and evaluation). When considering that the reliability of individual subscales of the most widely used instruments for measuring critical thinking ability are unstable at 0.21 through to 0.51, and 0.17 through to 0.74, respectively [11], construct validity may be deemed to be weak. It appears that subcategories such as interpretation, analysis, inference, and evaluation are applied mutually complementarily rather than being applied independently.

In conclusion, using IRT, the revised 19-item version of the CCTS showed relatively low levels of item difficulty and appropriate or high levels of discrimination. This revised CCTS has the advantage of enabling more convenient measurement of critical thinking skills than the 30-item CCTS [5] due to its improved reliability and validity. The levels of difficulty and discrimination of the revised CCTS-19 should be verified through retest and analysis so that it can be used to assess clinical critical thinking skills.

ORCID: Sujin Shin: http://orcid.org/0000-0001-7981-2893; Dukyoo Jung: http://orcid.org/0000-0002-0087-765x; Sungeun Kim: http://orcid.org/0000-0003-1195-0602

CONFLICT OF INTEREST

No potential conflict of interest relevant to this article exists.

ACKNOWLEDGMENTS

This work was supported by the research year grant of Soonchunhyang University (2014) [Fundref ID: 10.03039/501100002560].

SUPPLEMENTARY MATERIAL

Audio recording of the abstract.

REFERENCES

1. Shin K, Jung DY, Shin S, Kim MS. Critical thinking dispositions and skills of senior nursing students in associate, baccalaureate, and RN-to-BSN programs. J Nurs Educ. 2006 Jun;45:233-237.
2. Shin SJ, Jung D. Critical thinking in nursing science: a literature review. J Korean Acad Adult Nurs. 2009;21:117-128.
3. Kim MS, Park C, Kim KS. A study for developing critical thinking test (I): development of pilot test items. Seoul: Korean Institute for Curriculum and Evaluation; 2001.
4. Shin SJ, Yang E, Kong B, Jung D. Development and validation of a clinical critical thinking skills scale. Korean Med Educ Rev. 2012;14:102-108.
5. Hopfenbeck TN, Maul A. Examining evidence for the validity of PISA learning strategy scales based on student response processes. Int J Test. 2011;11:95-121.
6. Flora DB, Curran PJ. An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. Psychol Methods. 2004;9:466-491. http://dx.doi.org/10.1037/1082-989x.9.4.466
7. Seong T. Modern educational evaluation. Seoul: Hanjisa; 2014.
8. Murphy KR, Davidshofer CO. Psychological testing: principles and applications. 6th ed. Upper Saddle River (NJ): Pearson Education International; 2005.
9. Schmeiser CB, Welch CJ. Test development. In: Brennan RL, editor. Educational measurement. 4th ed. Westport (CT): Praeger Publishers; 2006.
10. Waltz CF, Strickland OL, Lenz ER. Measurement in nursing and health research. 4th ed. New York (NY): Springer Publishing Company; 2010.
11. Ku KY. Assessing students' critical thinking performance: urging for measurements using multi-response format. Think Skill Creat. 2009;4:70-76. http://dx.doi.org/10.1016/j.tsc.2009.02.001