Nurse Education Today 35 (2015) 864–874

Contents lists available at ScienceDirect

Nurse Education Today

journal homepage: www.elsevier.com/nedt

Review

Evaluation of tools used to measure critical thinking development in nursing and midwifery undergraduate students: A systematic review

Amanda G. Carter [a,*], Debra K. Creedy [b], Mary Sidebotham [a]

[a] School of Nursing and Midwifery, Centre for Health Practice Innovation, Menzies Health Institute Queensland, Griffith University, Brisbane, Australia
[b] Centre for Health Practice Innovation, Menzies Health Institute Queensland, Griffith University, Brisbane, Australia

ARTICLE INFO

Article history:
Accepted 25 February 2015

Keywords:
Critical thinking
Nursing
Midwifery
Measures
Scales
Evaluation

SUMMARY

Background: Well developed critical thinking skills are essential for nursing and midwifery practices. The development of students' higher-order cognitive abilities, such as critical thinking, is also well recognised in nursing and midwifery education. Measurement of critical thinking development is important to demonstrate change over time and effectiveness of teaching strategies.
Objective: To evaluate tools designed to measure critical thinking in nursing and midwifery undergraduate students.
Data Sources: The following six databases were searched and resulted in the retrieval of 1191 papers: CINAHL, Ovid Medline, ERIC, Informit, PsycINFO and Scopus.
Review Methods: After screening for inclusion, each paper was evaluated using the Critical Appraisal Skills Programme Tool. Thirty-four studies met the inclusion criteria and quality appraisal. Sixteen different tools that measure critical thinking were reviewed for reliability and validity and extent to which the domains of critical thinking were evident.
Results: Sixty percent of studies utilised one of four standardised commercially available measures of critical thinking. Reliability and validity were not consistently reported and there was a variation in reliability across studies that used the same measure. Of the remaining studies using different tools, there was also limited reporting of reliability making it difficult to assess internal consistency and potential applicability of measures across settings.
Conclusions: Discipline specific instruments to measure critical thinking in nursing and midwifery are required, specifically tools that measure the application of critical thinking to practise. Given that critical thinking development occurs over an extended period, measurement needs to be repeated and multiple methods of measurement used over time.

Introduction

The development of critical thinking (CT) skills has long been recognised as a priority in tertiary education. The landmark Delphi study by the American Philosophical Association (APA) produced an international expert consensus definition of critical thinking. Critical thinking is described as purposeful, self-regulatory judgement which results in interpretation, analysis, evaluation, and inference (Facione, 1990). Critical thinkers consider events or issues in a controlled, purposeful, focussed and conscious way (Mong-Chue, 2000).

Critical thinking is a crucial skill for nurses and midwives who, like other healthcare clinicians, need to effectively manage complex care situations in fast paced environments that demand increasing accountability (Mong-Chue, 2000; Muoni, 2012; Pucer et al., 2014). The processes of clinical decision-making and problem-solving require advanced CT skills (Muoni, 2012). CT is also essential for clinicians to critique and apply evidence, especially in situations where uncertainty regarding 'best practice' remains unclear (Scholes et al., 2012).

Although the development of students' higher order cognitive abilities is recognised as important in nursing and midwifery education, the measurement of these vital skills is inconsistent or neglected (Walsh and Seldomridge, 2006). The measurement of CT is important to identify deficits and developments in students' cognitive capacities as well as demonstrate the effectiveness of teaching strategies. The purpose of this systematic review was to evaluate tools used to measure CT development in nursing and midwifery undergraduate students.

Search Strategies Utilised

A search of major databases CINAHL, Ovid Medline, ERIC, Informit, PsycINFO and Scopus, was conducted in September 2014. The search was limited to English language articles published in peer reviewed

* Corresponding author at: School of Nursing and Midwifery, Griffith University, University Drive, Meadowbrook, Queensland 4131, Australia. Tel.: +61 7 33821535.
E-mail address: a.carter@griffith.edu.au (A.G. Carter).

http://dx.doi.org/10.1016/j.nedt.2015.02.023
0260-6917/

journals during 2001–2014. This period was chosen as the results of a Delphi study to define CT in nursing was published in 2000 (Scheffer and Rubenfeld, 2000). Scholarly work about CT in nursing would have further developed since that publication.

The inclusion criteria were original research studies that utilised experimental designs to assess CT development in undergraduate nursing and/or midwifery students. Papers were excluded if CT was not specifically measured on more than one occasion; the sample was post-graduate students, full text was not available in English, discussion papers that did not involve original research, or did not use an experimental design.

Five search terms were entered into the databases with the article title, abstract and body all searched. The search terms used were as follows:

1. "critical thinking" AND midwife*;
2. "critical thinking" AND midwife* AND measure*;
3. "critical thinking" AND midwife* AND evaluat*;
4. "critical thinking" AND students, nursing AND measure*; and
5. "critical thinking" AND students, nursing AND evaluat*.

The search was conducted sequentially using the search engines and search terms. An initial search, filtering for date, language and source of publication, identified 1191 papers. Once duplicates were excluded, each identified citation was reviewed using the inclusion and exclusion criteria and filtered through three screening levels i.e., (i) title screening; (ii) title and abstract screening; and (iii) full-text screening. Articles that were not relevant or did not meet the inclusion criteria were discarded. Finally 35 papers were included. No papers involving midwifery undergraduate students met the inclusion criteria and hence the samples in all of the papers are undergraduate nursing students.

Overview of Tools

Twenty-one (60%) of the 34 studies reviewed utilised one of four standardised commercially available measures of critical thinking. These were the California CT Disposition Inventory (10 studies), the California CT Skills Test (5 studies), the Watson–Glaser CT Appraisal (3 studies) and Health Services Reasoning Test (3 studies). Two studies used both the Californian CT Skills Test and California CT Disposition Inventory. All of these tools have reported psychometric reliability and validity allowing comparison across settings, disciplines, and time. Relatively few of the included studies (9 out of 21) undertook a reliability analysis of the tool for their current context. There were twelve other measurement tools utilised in the studies reviewed. See Table 1 for a comparison of tools employed in the studies reviewed.

Included studies were listed in a summary table (Table 2) during the search. The studies are presented in groups according to the tool utilised. After the initial search all articles identified in subsequent searches were checked against articles in the summary table and duplicates excluded. Each article was also entered into a reference management database (endnote) including the search term and engine used to locate each article. A quality appraisal process was performed using the Critical Appraisal Skills Programme (CASP) tool (CASP, 2013) and one article of poor quality was excluded. The excluded study is identified in the summary table. Following the quality appraisal process 34 papers were selected for review.

Results

All 34 studies measured CT skill development or change, either following completion of a specific educational intervention or an undergraduate nursing programme. Most studies were conducted in Western countries namely USA (n = 20), United Kingdom (n = 1), others were conducted in Taiwan (n = 4), Korea (n = 3), China (n = 2), Iran (n = 1), Hong Kong (n = 2), Turkey (n = 1), and Slovenia (n = 1).

Reliability, Validity and Factor Domains of the Tools

Reliability, validity and factor domains of the tools were examined. This included examination of previous and current reliability and validity testing. In respect to reliability, Facione and Facione (1992b) noted that a Kuder–Richardson (KR-20) range of .65 to .75 for this type of instrument is acceptable. Kaplan and Sacuzzo (1997) similarly reported that reliability estimates in the range of .70 to .80 are acceptable.

Factor Domains

In addition to developing a definition of CT, the APA also concluded that critical thinking comprised two dimensions; cognitive skills and disposition (Facione, 1990). Within the cognitive skills dimension, four sub-skills were defined; interpretation, analysis, evaluation, and inference. The disposition dimension was defined as truth-seeking, open-mindedness, analyticity, systematicity, self-confidence, inquisitiveness, and maturity of judgement (Facione and Facione, 1992a). Some scholars argued about the applicability of the universal definition of CT to the discipline of nursing. Scheffer and Rubenfeld (2000) conducted a Delphi study to develop a consensus definition of CT in nursing. A set of 17 consensus CT skills and habits of the mind were developed, many of which reflected Facione's (1990) earlier work with the addition of creativity, intuition and transforming knowledge (Scheffer and Rubenfeld, 2000). There has not been any published work on a definition of critical thinking for midwifery. The construct validity of the tools was assessed according to the dimensions and sub-skills of CT as outlined in the previous work of Facione (1990) and Scheffer and Rubenfeld (2000).

The California Critical Thinking Disposition Inventory (CCTDI) uses the APA consensus definition of critical thinking as the theoretical basis to measure the extent to which an individual possesses the attitudes of a critical thinker (Facione and Facione, 1992a). The domains assessed are as follows: open-mindedness, analyticity, cognitive, maturity, truth-seeking, systematicity, inquisitiveness, and self-confidence.

The CCTDI has a reported overall median alpha coefficient of .90 (Facione and Facione, 1994), demonstrating good reliability. Within the twelve studies that utilised the CCTDI only four (Atay and Karabacak, 2012; Shin et al., 2006; Stewart and Dempsey, 2005; Yu et al., 2012) tested reliability of the CCTDI. Two of the studies (Atay and Karabacak, 2012; Yu et al., 2012) reported reliability levels similar to those reported by Facione and Facione (1994) of .88 and .89. However, Stewart and Dempsey (2005) reported only marginal reliability with an alpha coefficient between .67 and .75. Shin et al. (2006) reported a much lower alpha coefficient of .53. These inconsistent results place some doubt on the reliability of this tool in different nursing education contexts.

The California Critical Thinking Skills Test (CCTST) was designed to measure critical thinking in college students (Facione and Facione, 1992b). The CCTST measures the ability of participants to draw conclusions in the areas of analysis, inference, evaluation, deductive and inductive reasoning (Facione and Facione, 1998). These skills relate to the APA consensus definition of critical thinking (Facione, 1990). The KR-20 estimate of internal consistency of the CCTST was r = .70 (Facione and Facione, 1998). Four of the seven studies that utilised the CCTST reported on reliability. Two studies reported low alpha coefficients of .62 (Beckie et al., 2001) and between .55 and .83 (Spelic et al., 2001). The CCTST was used to track development of CT in students undertaking different study pathways (Spelic et al., 2001). Some concerns were expressed with the internal consistency of the CCTST across the different cohorts. The total score α for the RN-BSN group was very low (alpha = .31) compared to the traditional and accelerated pathways cohorts (alpha = .66). Spelic et al. (2001) suggested that the reliability of tools with few items and involving a timed test administration is low. The CCTST comprises 34 items, and Spelic et al. (2001) found that on several items all students scored the same. When these items were removed the α level for the 30 items was .62. This limitation highlights the value of using multiple measures in the assessment of CT.