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

Author

Carter, Amanda G, Creedy, Debra K, Sidebotham, Mary

Published

2015

Journal Title

Nurse Education Today

Version

Accepted Manuscript (AM)

DOI

10.1016/j.nedt.2015.02.023

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

Word Count 4494 (without references and tables)

Authors:
Amanda G. Carter RM BHealthSc MMid
School of Nursing and Midwifery
Griffith University, Brisbane, Australia.

Debra K. Creedy RN PhD
Professor, Centre for Health Practice Innovation
Griffith Health Institute
Griffith University, Brisbane, Australia.

Mary Sidebotham RM PhD
School of Nursing and Midwifery
Griffith University, Brisbane, Australia.

Corresponding Author
Amanda G. Carter
School of Nursing and Midwifery
Griffith University
University Drive
Meadowbrook. Queensland 4131, Australia
Ph +61 7 33821535
a.carter@griffith.edu.au

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

Abstract

Background: Well developed critical thinking skills are essential for nursing and midwifery practice. 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; CINAHL, Ovid Medline, ERIC, Informit, PsycINFO and Scopus were searched and resulted in the retrieval of 1,191 papers.

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 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 practice. Given that critical thinking development occurs over an extended period, measurement needs to be repeated and multiple methods of measurement used over time.

Key words: critical thinking, nursing, midwifery, measures, scales, evaluation

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 judgment 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, Trobec, & Žvanut, 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 & 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 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 & 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:

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*
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 1,191 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 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.

Table 1: Description of Tools/Methods to measure critical thinking font in the table differs from the text. ? make them all the same

Name of Instrument/ Author/ Year Developed	Aim of tool	Number of Items/ format	Psychometric Testing	Scores	Time to Complete	Factor Domains Measured
The California Critical Thinking Disposition Inventory (CCTDI) / Facione & Facione/ 1992	Measure the extent to which an individual possesses the attitudes of a critical thinker. Designed for use by the general adult population	75 Likert items, "agree-disagree" scale, student's self report	Cronbach's alpha .90 for the overall instrument and .71 to .80 for the seven subscales	Maximum score of 60 in each domain. Negative disposition is a score below 30. The total maximum score is 420 points. Scores > 350 indicate a high CT disposition. Scores less < 280 indicate paucity of CT	20-30 mins	Open-mindedness, analyticity, cognitive, maturity, truth-seeking, systematicity, inquisitiveness, and self-confidence.
California Critical Thinking Skills Test (CCTST)/ Facione & Facione/ 1992	Designed for assessment of entry or exit level CT skills of various groups of college students and for evaluation of learning outcomes of various curricular programs.	34 Multiple choice items uses a generic scenario requiring an accurate and complete interpretation of the question	The Kuder-Richardson (KR-20) estimate of internal consistency of the CCTST is reported in the test manual to be r = .70	The maximum total score is 34. A score of ≥24 indicates very strong CT skills. A score 13-23 indicates a mid-range skill level Scores of ≤ 12 indicate fundamental weaknesses in CT skills.	45-50 mins	Analysis, inference, evaluation, deductive and inductive reasoning
Health Sciences Reasoning Test (HSRT) /Facione, Facione, & Winterhalter/	Adaptation of the CCTST specifically designed for use by health sciences students	33 multiple choice questions uses a health related scenario requiring an	Internal consistency .77 to .84. overall internal consistency value of .81 with Kuder-Richardson formula 20, and an overall .81 reliability coefficient	Total score reflects overall CT skills. Maximum score is 33. Scores of 25 or above represent strong CT skills,	30-50 mins	Analysis, inference, evaluation, inductive reasoning and deductive reasoning

2010	and professionals to assess their CT and clinical reasoning skills.	accurate and complete interpretation of the question		scores from 15 to 24 are considered mid-range and represent competence in CT skills in most situations, and scores of 14 or below represent fundamental weaknesses in CT skills		
The Watson-Glaser Critical Thinking Appraisal (WGCTA) / Watson & Glaser/ originally developed in 1925, most recent revision 2012	Measures both logical and creative components of CT and assesses CT ability in individuals with at least a ninth grade education	40 multiple choice items answering scenario based questions	Reliability reported to be > .8. Using the Spearmen-Brown formula, reliability for the total score of the WGCTA was established at .77. This is consistent with the split-half reliability coefficients, ranging from .76 to .85	Maximum score is 80	40-50 mins	Inference, recognition of assumptions, deduction, interpretation and evaluation of arguments
Think aloud analytic framework / Daly/ 2001	Analyse qualitative data to synthesise conception of CT	A scale of argument/epistemological complexity is used to assess videotaped client simulation	Not stated	Scores range from 1-4	No time commitment by student. Uses learning activities integrated into the course	Structural components of differentiation and integration in reasoning, situation modelling and argument and evidential structure.
Critical Thinking Ability Scale (CTAS) for	Assess dimensions of CT of college	20 items measured using a Likert scale 1	Cronbach's alpha was found to be .74 (Park, 1999)	Total scores have a possible range from 5 to 100, with higher	Not stated	Intellectual curiosity, healthy skepticism, intellectual integrity,

College Students/ Park/1999	students	= absolutely do not agree to 5 = absolutely agree		score indicating stronger CT ability		prudence, and objectivity
Critical Thinking Disposition Scale for Nursing Students (CTDS) /Park & Kim /2009 (Korean version only)	Assess of CT disposition in Korean nurses	35 items assessed a 5-point Likert scale. Student self-report	Cronbach's alpha = .78 (Park & Kim, 2009)	The total score ranges from 35 to 175, with a higher score indicating a higher level of critical thinking disposition	No stated	Intellectual integrity, creativity, challenge, open-mindedness, prudence, objectivity, truth seeking, inquisitiveness,
Critical Thinking Process Test (CTPT)/ Educational Resources Inc./ 1999	Developed specifically for nursing students. Focus on critical thinking process skills within a nursing environment, not level of nursing content knowledge	50 item multiple choice	The average reliability coefficient was .93 with demonstrated evidence of content and diagnostic validity (Anderson et al, 2000).	Not stated	60 minutes	Assesses 4 aspects of the critical thinking process: listening, writing, speaking, and reading, and 5 levels of abstract thinking: prioritizing, inferential reasoning, goal setting, application of knowledge, and evaluation of predicted outcomes.
Think aloud protocol / Morey, 2002	Provide a valid source of qualitative data on thinking and thought processes	A rating tool and rubric using a 4 point Likert scale for eight cognitive processes, level of critical thinking, and for	Two faculty rated the think-aloud scenario responses with 97.9 to 100 percent rater agreement.	Not provided	No time commitment by student as uses learning activities integrated into the course	Collect, review, relate, interpret, infer, diagnosis, act, and evaluate

		accuracy of nursing diagnosis, conclusions, and evaluation.				
N3 case report accreditation form /Taiwan Nurses Association / no date available	Not stated	45 criteria (including 36 strengths and 9 weaknesses	Inter-rater reliability = .893, internal consistency of KR-20 = .79 and test-retest reliability of .32 ($p<0.01$).	Total scores ranged from 0-45.	No time commitment by student. Uses learning activities integrated into the course	Constructed on the basis of the nursing process. Critical inquiry points are listed under each step of the nursing process
Discussion board analysis/ Pucer Trobec & Žvanut / 2014	Analyse discussion board posts for evidence of CT	Discussion posts examined against six elements of critical thinking	Not stated	Not stated	60 minutes	Analysis, inference, interpretation, explanation, evaluation, and self-regulation.
Critical Thinking scale (CTS) / Cheng, Wang, Wu, & Hwang, / 1996	Not provided	60 item multiple choice questions. Participants choose one correct answer from either one in five or dichotomous response sets according to the item situations.	CTS demonstrated adequate reliability (internal consistency as well as split half reliability) and convergent as well as known group validity.	The higher scores indicate better CT skills	Not provided	Inference, recognition of assumptions, deduction, interpretation, and evaluation of argument.
Critical Thinking Assessment (CTA) / Assessment	Determine students' overall performance on specified CT skills	40 generic multiple choice questions	CTA has a global alpha of .69 and a standardized item alpha of .70 for all 40 items in first-time	Maximum score of 40	Not provided	Interpretation, analysis, evaluation, inference, explanation, self-

Technologies Institute / 2001	determined to be necessary for success in an academic program for nursing study.		examinees (ATI, 2001).			regulation
Blooms Taxonomy / Jones, 2008	Assess student's developed nursing care plans for evidence of critical thinking	Using nursing care plans	Not provided	Not provided	No time commitment by student as uses learning activities integrated into the course	Knowledge, comprehension, application, analysis, synthesis, evaluation
Concept map scoring / Daley Shaw, Balistrieri, Glasenapp, Piacentine / 1999	Assess student's ability to develop concept maps that reflect CT used in the nursing process.	Using concept maps	Inter-rater reliability was performed with two assessors in the pilot study and the percentage of agreement of the independent scores was 85%. Content validity establish by Daley et al (1999).	Not provided	No time commitment by student as uses learning activities integrated into the course	Meaningful, valid and significant
Critical Thinking Scale (CTSM) / McMaster University / 2002	Not provided	10 items. Each item is scored on a six-point Likert scale of 1 to 6, with 1 corresponding to "never" and 6 to "always". .	Cronbach's alpha coefficient .93 and two-week test-retest reliability coefficient was .92.	Total scores range from 10 to 60 with higher scores indicating higher level of CT competency	Not provided	Not stated

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.

Table 2: Articles that met inclusion and quality criteria

Author, year and location	Design/Intervention	Participants	Results	Reliability and validity assessment	Quality Appraisal using CASP
Californian Critical Thinking Disposition Inventory (CCTDI)					
Atay, & Karabaca (2012). Turkey	Pre- post-test control group design testing effects of using concept plans	80 freshman and sophomore nursing students	Statistically significant increase in CT scores for experimental group.	Cronbach's alpha for the was .88.	Include
Shin, Lee, Ha, & Kim (2006) Korea	Longitudinal study using CCTDI each year for 4 years	60 nursing students commenced on study, 32 completed all four surveys	Statistically significant improvement in CT disposition	Cronbach's alpha for the CCTDI was .59 in Yr 1, .53 for Yr 2, .66 for Yr 3, and .73 for Yr 4. Significantly lower than overall median alpha coefficient of .90 reported by Facione (1994)	Include
Tiwari, Avery, & Lai (2006). Hong Kong	Experimental design, pre-post test testing the effects of PBL. 4 time points tested	79 1st year nursing students.	Significantly greater improvement in CT scores for experimental group	No reporting of reliability of CCTDI for this study.	Include
Evans & Bendel, (2004). United States	Quasi-experimental, non-equivalent control group design testing narrative pedagogy	114 undergraduate nursing students,	No significant differences in CT scores between control and experimental groups	No reporting of reliability of CCTDI for this study.	Include
Wood & Toronto (2012) USA	Experimental study testing the effects of human patient simulation	85 2nd year nursing students	Higher mean post-test total scores compared with pre-test total scores in experimental group students.	No reporting of reliability of CCTDI for this study.	Include
Stewart & Dempsey (2005). USA	Longitudinal study, at 5 time-points testing effects of whole program	55 nursing students recruited, 34 students completed all surveys	Subscale and total scores did not significantly increase throughout the program.	Cronbach's alpha for the CCDDI was calculated at each phase: Sophomore semester 2 =	. Include

				.71. Junior semester 1 = .77 Junior semester 2 = .76 Senior semester 1 = .67 Senior semester 2 = .75	
Yeh & Chen (2005). Taiwan	A pre- and post-test quasi-experimental research design testing the effects of a CT lecture and interactive videodisc system	126 RN-BN students	Statistically significant differences between pre and post-test overall scores	No reporting of reliability of CCDTI for this study.	Include
Yu, Zhang, Xu, Wu & Wang (2012). China	Crossover experimental study testing the effects of PBL	76 2nd year nursing students.	Statistical improvement in overall CTDI scores following PBL	For this study the overall Cronbach's alpha was .8999	Include
Dehkordi, & Heydarnejad, (2008). USA	Quasi-experimental design testing the effects of PBL	40 2nd year nursing students participated.	Statistical improvement in CTDI scores following PBL	No reporting of reliability of CCDTI for this study	Include
Zadeh, Khajeali, Khalkhali, & Mohammadpour (2014). Iran	Quasi-experimental study testing the effects of an evidence based nursing course	48 3rd year nursing students	CCTDI scores were significantly higher following the intervention	No reporting of reliability of CCDTI.	Include
Californian Critical Thinking Test (CCTST)					
Chau, et al (2001). Hong Kong	Pre-test/post-test design testing the effects of 4 vignettes.	101 1st and 2nd year nursing students recruited of 83 completed both pre and post-tests.	No statistical difference in pre and post test scores.	KR-20 of the CCTST was .74 and subscales ranged from .30 to .61.	Include
Beckie,	A pre-post test, non-	183 BN students	Cohort 1 received the new	Cronbach alpha on CCTST	Include

Lowry, & Barnett, (2001). United States	equivalent control group design. Experimental group experienced new curriculum	consisted of 3 cohorts of students, 1 control cohort and 2 cohorts that experienced the new curriculum	curriculum, achieved significantly higher CT scores than controls. Cohort 3, the 2nd class to experience the revised curriculum, failed to demonstrate improved CT scores and reported some decreases.	ranged from .55 to .83. Internal consistency of tool low and varied across tests.	
Spelic, et al.,(2001). United States	Longitudinal study testing effects of different pathways	136 students in 3 undergraduate pathways, traditional, accelerated and RN-BSN	Statistically significant increase in CT scores for all pathways	The CCTST has 34 items. No demonstrated variance (all students scored the same) on some items, α level therefore computed on less than 30 items.	Include
Wheeler, & Collins, (2003) United States	Quasi-experimental design. Testing the effects of concept mapping compared to traditional nursing care plans.	A convenience sample (n = 76)	Significant difference between pre – post test scores for both groups. No difference found between experimental and control groups.	No reporting of reliability of CCTST for this study.	Include
Yuan, Kunaviktikul, Klunklin, & Williams, (2008). China	A quasi-experimental, two-group pre–post test design testing the effects of PBL	All 46 Year 2 nursing students	PBL students had significantly greater improvements on overall CCTST	KR20 for the CCTST-A was .80 for the total scale and between .60-.78 for subscales.	Include
Californian Critical Thinking Skills Test (CCTST) & Californian Critical Thinking Disposition Inventory (CCTDI)					
Ravert, (2008). States	Pre-post test design testing effects of human patient simulation	30 1st year students	No differences in CT scores	No reporting of reliability of the CCTST or CCTDI for this study.	Include
Naber & Wyatt, (2014) United States	Experimental, pre–post test design testing effects of reflective writing	70 4th semester nursing students	The experimental group's total CCTST and CCTDI scores did not increase significantly following the intervention.	No reporting of reliability of CCTST or CCTDI scale for this study.	Include

Health Sciences Reasoning Test (HSRT)					
Sullivan-Mann, Perron, & Fellner (2009). United States	Mixed-model experimental design, testing effects of multiple simulation	53 nursing students from the medical-surgical course	Statistically significant increase in CT scores for experimental group.	Reliability of the HRST not reported for this study.	Include
Shinnick,. & Woo, (2013). United States	One-group, quasi-experimental, pre-post test design. Tested the effects on one human patient simulation	A convenience sample of 154, 3rd or 4th year nursing students	Following HPS there were no statistically significant gains in CT, with some decrease in scores (not statistically significant).	No reporting of reliability of HSRT for this study.	Include
Goodstone et al, (2013). USA	A two-group quasi-experimental pre-port test design testing the effects of high fidelity patient simulation (HFPS) compared to case study	42 1st semester associate degree nursing students. Allocated to two groups, HFPS, and case study group,	There was a significant increase in the HSRT scores for the case study group (p=0.003) but not for the HFPS group.	.No reporting of reliability of HSRT for this study	Include
Watson-Glaser Critical Thinking Appraisal (WGCTA)					
L'Eplattenier (2001). United States	Longitudinal study testing 4 times over 3 year undergraduate program	83 nursing students	No change in CT scores as student progressed through the program.	No reporting of reliability of WGCTA for this study.	Include
Brown, Alverson, & Pepa (2001). United States	Longitudinal study, testing at the beginning and end of degree. Testing different pathways and length of program	Convenience sample (n = 123) of three groups of baccalaureate nursing students: traditional, RN-BSN, and accelerated.	A significance difference found between pre- and post WGCTA scores for traditional students (p=0.007) and RN-BSN (p=0.029), with no difference for accelerated students.	Reliability for the total score of the WGCTA was established at .77 (using Spearmen-Brown formula). Consistent with the split-half reliability coefficients (.69 to .85), reported by Watson and Glaser	Include
Watson-Glaser Critical Thinking Appraisal (WGCTA) and Think Aloud Analytical Framework					
Daly,	A longitudinal multi-	43 nursing students	No statistical difference in	No reporting of reliability of	Include

(2001). United Kingdom	method design with triangulation.	completed WGCTA. 12 students completed think aloud analytical framework	WGCTA scores. Little evidence of CT demonstrated in think aloud analytical framework	WGCTA for this study. No discussion of reliability or validity of think aloud analytical framework. Not clear whether the think aloud tool was validated or reviewed by experts and inter-rater reliability was not discussed	
Critical Thinking Ability Scale (CTAS) for College Students					
Choi, Lindquist, & Song, (2014). Korea	Non-equivalent control group pre–post test design testing effects of PBL.	90 1st year nursing students	No significant differences in CT scores between control and experimental groups	Cronbach's alpha was .71 which is consistent with the reported .74 by Park (1999). Not available in English	Include
Critical Thinking Disposition Scale for Nursing Students (CTDS)					
Jun, Lee, Park, Chang & Kim (2013). South Korea	Quasi-experimental study testing effects of 5E learning cycle model with PBL	161 1st year nursing students	Statistically significant increase in CT scores for experimental group.	Cronbach's alpha was .81. CTDS not available in English, 20 point self report Likert scale measures disposition as a proxy for CT skills.	Include
Critical Thinking Process Test (CTPT)					
DeSimone, (2006). United States	Experimental design testing effects of accelerated program	38 nursing students undertaking an accelerated program (12 months in length)	Increase in CT scores not significantly different	Average reliability coefficient was .93.	Include
Critical Thinking Process Test + Think Aloud Protocol					
Morey, (2012). United States	An experimental design testing an online animated pedagogical agent.	45 associate degree nursing students in their final semester	No differences in CT for either tool	No reporting of reliability of CTPT. Two faculty rated the think-aloud scenario responses with 97.9 to 100	Include

				percent rater agreement. Limited information provided regarding the think aloud protocol.	
N3 Case Report Accreditation Form					
Chen, & Lin, (2001) Taiwan	Quasi- experimental design with pre-post test testing effects of a research course	168 1st year nursing students.	Experimental group reported significantly higher CT scores than control group	No reporting of reliability of the N3 case report form. Unclear whether tool measured students' ability to critique an article rather than CT abilities..	Include
Discussion Board Analysis					
Pucer, Trobec, & Žvanut, (2014) Slovenia	Quasi-experiment study testing the effects of an ICT program which presented scenarios that mirror clinical situations.	45 1st year nursing students	Qualitative analysis of the discussion boards showed a significant improvement in % of posts for which the opinions and conclusions of the participants were justified with valid arguments.	No reporting of tool reliability. No discussion regarding development of tools, expert review process or psychometric testing of the tool	Include
Critical Thinking Scale (CTS)					
Lee et al (2013) Taiwan	Longitudinal study, measuring at 4 time-points testing the effects of concept mapping	A convenience sample of 95 students,	Both control and experimental groups had higher initial CT scores that tended to decrease over time.	No reporting of reliability of CT scale for this study.	Include
Critical Thinking Assessment (CTA)					
Mann, (2012). USA	Experimental, pre-post-test, mixed method design testing the effects of grand rounds	21 2nd year nursing students.	No significant difference between CT scores for the two groups. In the control group, students' scores indicated a decrease CT scores.	No reporting of reliability of CTA for this study.	Include
Blooms Taxonomy					

Jones,(2008).USA	A quasi-experimental, pre-post test study testing the effects of PBL	60 2nd year nursing students.	Intervention group demonstrated a higher significant increase in CT compared to the control group .	No reporting of reliability. Unclear whether the tool was validated or reviewed by experts. Blooms taxonomy used to develop the tool, but no attempt to relate this to the recognised definitions of CT	Include
Concept Map Scoring					
Abel. & Freeze, (2006) USA	Longitudinal study measurement over 4 timepoints testing the effects of concept mapping	28 associate degree nursing students	There was a significant increase in mean scores of the first concept map to the average mean score of the last two maps (p=0.05).	No reporting of reliability of tool. Limited information about scoring criteria, needed more information how this score relates to critical thinking	Include
Critical Thinking Likert Scale (CTLS)					
Stevens, Brenner & Brenner (2009) USA	Pre-post test experimental design testing the PALS learning approach	15 nursing students	Increase in scores on CTLS but no statistical analysis performed.	No reporting of reliability of CTLS for this study or previously.	Exclude due to lack of statistical analysis and reporting of results.
Critical Thinking Scale (CTSM)					
Tseng, et al (2011). Taiwan	A quasi-experimental design measurement over 3 time-points testing the effects of PBL.	120 RN students.	The CTS scores were significantly higher in the experimental group	Cronbach's alpha coefficient of the CTS was .94. Limited information regarding the CTS tool and how it measured CT.	Include

Results

All 34 studies measured CT skill development or change, either following completion of a specific educational intervention or an undergraduate nursing program. 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 judgment (Facione & 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 & 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 & Facione, 1992a). The domains assessed are: open-mindedness, analyticity, cognitive, maturity, truth-seeking, systematicity, inquisitiveness, and self-confidence.

The CCTDI has a reported overall median alpha coefficient of .90 (Facione, 1994), demonstrating good reliability. Within the twelve studies that utilised the CCTDI only four (Atay & Karabacak, 2012; Shin et al., 2006; Stewart & Dempsey, 2005; Yu et al., 2012) tested reliability of the CCTDI. Two of the studies (Atay et al., 2012; Yu et al., 2012) reported reliability levels similar to those reported by Facione (1994) of .88 and .89. However, Stewart and Dempsey (2005) reported only marginal reliability with an alpha coefficient between .67 and .75. Shin (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, 1992b). The CCTST measures the ability of participants to draw conclusions in the areas of analysis, inference, evaluation, deductive and inductive reasoning. (Facione & 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 & 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 30 items was .62. This limitation highlights the value of using multiple measures in the assessment of CT.

The second study using the CCTST demonstrated inconsistent results (Beckie, et al., 2001). Two cohorts of nursing students in a new curriculum focussing on CT skills completed the CCTST over three time-points. The first group experienced significantly improved CT scores from baseline but scores of the second group revealed decreased CT scores. This variation in results across the two cohorts undertaking the same curriculum places doubt on the reliability of this tool.

The other two studies that tested the reliability of the CCTST (Chau et al., 2001; Yuan et al., 2008) reported similar results to Facione and Facione (1998). The differences in findings

between these four studies may indicate that the CCTST does not consistently measure CT in nursing practice across different settings.

The HSRT is a commercially available, recent adaptation of the CCTST specifically designed for health sciences students and professionals to assess their CT and clinical reasoning skills (Goodstone et al, 2013). Similar to the CCTST the HSRT uses the sub-skills identified within the APA consensus definition of critical thinking. The HRST is considered a reliable and valid measure of critical thinking for entry level nursing students with a KR 20 of .81 (Facione, Facione & Winterhalter, 2010). The three studies that used this tool all tested the effects of simulation on CT but none reported reliability (Sullivan-Mann, et al, 2009; Shinnick & Woo, 2013; Goodstone et al, 2013). One study (Sullivan-Mann, et al, 2009) reported an increase in student's CT skills following simulation but the other two studies (Goodstone et al,2013; Shinnick & Woo, 2013) reported no statistical increase, with decrease in scores in one study. These inconsistent results could indicate the HSRT is not a reliable tool across diverse settings and populations.

The WGCTA, originally developed in the 1920's, measures both logical and creative components of CT and assesses CT ability in individuals with at least a ninth grade education (Watson & Glaser, 1980). The test comprises 80 proposed arguments related to 25 statements that include problems, arguments, and interpretations. On completion a total score is produced based on the assessment of five critical thinking skills: inference, recognition of assumptions, deduction, interpretation and evaluation of arguments, which align to the CT sub-skills defined by Facione (1990). The WGCTA measures the underlying constructs of classical logic and general reasoning skills rather than application of CT skills (Walsh & Seldomridge, 2006). Only the study by Brown et al. (2001) reported an alpha coefficient of .77. This is consistent with the split-half reliability coefficients of .69 to .85 reported by Watson and Glaser (1980). The three studies that used the WGCTA were all conducted in the USA and used a longitudinal design to detect change in CT across different undergraduate nursing degrees (L'Eplattenier, 2001; Brown et al, 2001; Daly, 2001). Two of the studies (L'Eplattenier, 2001; Daly, 2001) found no change in CT scores whereas Brown et al (2001) reported increases in CT scores of students undertaking traditional and RN-BSN pathways but no change for students in the accelerated pathway. These inconsistencies in findings may support claims that the constructs within the WGCTA are not suited to measure CT skills in the nursing discipline (Walsh & Seldomridge, 2006).

Of the twelve non-standardised tools utilised to measure critical thinking in this review, only four tested reliability. The Critical Thinking Ability Scale (CTAS) for College Students has a

reported Cronbach's alpha of .74 (Park, 1999). The CTAS was used by Choi et al (2014) to measure the effect of problem based learning (PBL) on CT and had a reported Cronbach's alpha of .71. Although the aim was to measure changes in students' CT abilities, the CTAS is a self-report tool that assesses the domains of; intellectual curiosity, healthy skepticism, intellectual integrity, prudence, and objectivity, which relate more to CT disposition rather than skills.

The Critical Thinking Disposition Scale (CTDS) for Nursing Students developed by Park & Kim (2009) has a reported Cronbach's alpha of .78. Jun et al. (2013) used the CTDS to measure critical thinking development in 161 nursing students, and reported a Cronbach's alpha of .81. The CTDS uses the concepts of intellectual integrity, creativity, challenge, open-mindedness, prudence, objectivity, truth seeking, inquisitiveness, which directly relate to dispositional characteristics identified by both Facione, (1990) and Scheffer & Rubenfeld (2000). This tool is not available in English which limits use in other settings. Similar to the CCTDI, the CTDS only measures CT disposition not the application of these skills in practice.

The N3 case report accreditation form developed by the Taiwan Nurses Association was used to assess students' CT abilities in the critique of case study reports (Chen & Lin, 2003). Testing of this tool resulted in good inter-rater reliability = .89 (Pearson r), internal consistency of KR-20 = .79, but low test-retest reliability of .32 after a 16 week interval. However, the construct validity of this tool is questionable. The criteria of the tool do not reflect any of the CT constructs. Instead the tool was constructed on the basis of the nursing process with critical inquiry points listed under each step of the nursing process (Chen & Lin, 2003). The study tested the effects of a research course, and found significantly higher CT scores in students who undertook the course. However, it was unclear whether the tool measured students' abilities to critique an article rather than their CT abilities.

The Critical Thinking Process Test (CTPT), a commercial tool developed by Educational Resources, has a reported reliability coefficient of.93 (Anderson et al, 2000). The CTPT measured CT development in two studies but neither reported on reliability (DeSimone, 2006; Morey, 2012). The CTPT assesses four aspects of the critical thinking process; listening, writing, speaking, and reading, and five levels of abstract thinking; prioritizing, inferential reasoning, goal setting, application of knowledge, and evaluation of predicted outcomes. Several concepts partially relate to elements of the recognised definition of CT. This tool is expensive to administer and not widely used (Fountain, 2011).

The Critical Thinking Scale (CTSM) developed by McMaster University assesses the effects of PBL and concept mapping on CT (Tseng et al., 2011). The reported Cronbach's coefficient of .94 (Tseng et al., 2011), was replicated in another study which reported .93 (Chou, Jian, Tseng & Ko, 2014). The concepts of inference, recognition of assumptions, deduction, interpretation, and evaluation of argument reflect the critical thinking sub-skills identified by Facione (1990). The CTSM is a student self-report test but may not measure CT in practice.

A validated concept map scoring criteria was used to measure CT development over a one year period (Abel & Freeze, 2006). Inter-rater reliability with two assessors found an 85% level of agreement (Abel & Freeze, 2006). The authors stated that content validity had previously been established, and no further testing of internal consistency was performed. The scoring criteria were: 1) meaningful relationships between two concepts indicated by a connecting line; 2) hierarchy shows a general to specific approach; 3) cross-links show meaningful connections between one segment of the hierarchy; and 4) examples describe specific instances of a concept (Lawson, 2012). It was unclear how the scoring criteria related to the dimensions of CT. The study demonstrated increases in students' concept map scores as they progressed through the curriculum, but it is uncertain whether this increase was representative of increases in critical thinking or simply improved competence in concept mapping.

The Critical Thinking Scale (CTS) assesses CT through the concepts of inference, recognition of assumptions, deduction, interpretation, and evaluation of argument (Lee et al., 2013). These concepts match those suggested within the two recognised definitions of CT. In a study examining the effects of concept mapping on CT skills, Lee et al., (2013) reported that previous reliability testing convergent as well as known group validity was conducted by the developer of the tool Cheng et al (1996). No further testing of the reliability of the tool was conducted by Lee et al. Using a longitudinal design, students' CT scores were compared between those exposed to one semester of teaching on concept mapping with a control group (Lee et al, 2013). Initial increases in CT scores were found in both groups but decreased over time. These findings infer the teaching methodologies were not effective but also may indicate the CTS is not reliable in measuring changes in CT over time.

The Critical Thinking Assessment (CTA) tool was used to evaluate the effects of a grand round education strategy on CT (Mann, 2012). The CTA has a reported alpha of .69 and a standardized item alpha of .70 in first-time examinees (Assessment Technologies Institute, 2001). No reliability testing was performed by Mann (2012). The CTA uses 40 multiple

choice questions based on the domains of interpretation, analysis, evaluation, inference, explanation and self-regulation (ATI, 2003). Four of these domains (interpretation, analysis, evaluation and inference) directly relate to the recognised domains of CT. There were no differences in the CT scores in the control or experimental groups, with a decrease in scores in the control group (Mann, 2012). The unexpected decrease in CT scores could be due to the very small sample size of 21, with only 4 students in the control group.

Four of the twelve non-standardised tools were newly developed with the specific purpose of measuring critical thinking in action (Daly, 2001; Jones, 2008; Morey, 2012; Pucer et al., 2014).The studies utilised practice-based teaching, learning, and assessment activities to measure CT which not only presents opportunities to evaluate the application of CT but also reduces survey and response burden as the activities are embedded in student learning. However, none of these studies reported reliability of these newly developed tools.

Pucer et al. (2014) used a discussion board tool to analyse student's postings according to identified core key elements of critical thinking (as defined by Facione, 1990). A significant improvement in the percentage of posts where the opinions and conclusions of participants were justified with valid arguments was reported (Pucer et al., 2014). However, limited information was presented on the development of the tool, process of expert review and validation, or inter-rater reliability.

The effect of PBL on students' CT development was measured by grading nursing care plans over a semester (Jones, 2008). The grading system was based on the six levels of Blooms taxonomy of cognitive learning and were described as; comprehending information, organising ideas, and evaluating information and actions. Students who experienced the PBL educational intervention reported higher CT scores. It was not clear however, whether the tool was validated or reviewed by experts. Although Blooms taxonomy was used as the basis of the tool, there did not seem to be any attempt to relate the grading domains to the recognised definitional elements of CT (Facione, 1990; Scheffer & Rubenfeld, 2000).

In an attempt to establish concurrent validity Morey (2012) used both a newly developed qualitative tool based on a 'think aloud protocol', and a standardised tool (CTPT) to measure the effects of an animated pedagogical agent on critical thinking. The think aloud protocol used elements of the nursing process to assess students' thinking in solving a clinical scenario (Morey, 2012). The elements of collect, review, relate, interpret, infer, diagnosis, act, and evaluate did not align directly with the recognise definitions of CT. Both groups displayed significant improvements in CT levels and correct conclusions from baseline to post-intervention on the think-aloud protocol, but only the pedagogical agent group had a

significant result for “evaluation”. These mixed results may indicate the difficulty in measuring CT development in a standardised exam format. Reliability testing and construct validity of the think aloud were not reported, therefore results must be viewed with caution.

Daly (2001) also compared the use of a newly developed think-aloud analytic framework and a standardised tool (WGCTA) to measure CT development over an 18 month period. No statistical improvement in the WGCTA scores was found. The think aloud qualitative assessment demonstrated consistent evidence of reasoning that reflected an “enduring absolutist epistemology” but portrayed little evidence of CT (Daly 2001). The authors explained that reasoning of this nature usually involves a single theory structured argument which is contradictory to the principles of CT (Daly, 2001). Although both tools indicated similar results, no reliability testing was conducted. The constructs of this new tool were described as differentiation and integration in reasoning, situation modelling and argument and evidential structure (Daly, 2001), which do not incorporate the recognised definitional elements of CT (Facione, 1990; Scheffer & Rubenfeld, 2000).

Discussion

This review included studies from 9 different countries using 16 different tools. This section discusses the findings in relation to the reliability, validity and factor domains of the standardised tools and then examines the non-standardised tools.

The reliability of tools used to measure CT in nursing practice was not reported consistently and varied considerably. Only two authors of new tools reported on internal stability using a test-retest, and at best, split-half reliability for internal consistency was reported. The review included four commercially available tools and this cost may limit their use for routine evaluation of classroom teaching effectiveness. The CCTDI and the CCTST had reported reliability ranging from .31 to .89 and some authors using these tools did not test reliability for their specific context. The CCTDI measures students’ self-report CT disposition and does not measure the development of CT skills. Relying on student self-report may be affected by recall bias and a socially desirable response set (Tiwari et al., 2006). The act of critical thinking involves both skills and habit of the mind (Scheffer & Rubenfeld, 2000). The CCTDI only measures the habits of the mind. For a complete assessment of student’s critical thinking both skills and disposition need to be measured, and the CCTST should be used in conjunction with the CCTDI (Insight Assessment, 2013).

A lack of congruence between items in the CCTST and the CCTDI could account for inconsistencies in reliability. Although the cognitive skills underlying the framework for the

CCTST and the CCTDI were identified as important to the practice of nursing (Stone, Davidson, Evans, & Hansen, 2001), the same study found less agreement on whether the items reflected CT skills required of nurses. Inconsistent results across studies have prompted questions related to the reliability of the CCTDI to measure dispositional attitudes (Walsh & Seldomridge, 2006), and the lack of stability of the instrument (Walsh & Hardy, 1997; Kakai, 2003).

Limited reporting of tool reliability makes it difficult to assess their applicability in the nursing and midwifery contexts. Concern could also be justified over the focus of existing tools (especially standardised tools) on the measurement of formal logic and general thinking skills, rather than the application of CT in practice (Seldomridge & Walsh, 2006).

Four new tools that measure the application of CT skills in nursing in practice were reviewed. However, none of these new tools were tested for reliability. When the domains were compared to the recognised definition of CT, construct validity was only established for one tool (Jones, 2008). None of the studies conducted a factor analysis to establish validity. In the development of the new tools, items were drawn from concepts thought to be useful but no testing was conducted to confirm this. Therefore, further research with large samples, factor analysis, and testing of different forms of reliability and validity, are required before implementing these tools into practice.

CT is also considered to be a multidimensional concept, and a single test in a multiple choice format may be inadequate to accurately detect change in development. There is a need to ensure that measures of CT development address the complexity of practice and are adaptive to the nursing and midwifery environments (Rubenfeld & Scheffer, 2006). A mixed method approach and triangulation of findings may provide greater validity, reliability, and insight into CT development.

Conclusion

There was limited reporting of the reliability of tools in the included studies. Overall there was relatively little emphasis placed on validity of newly developed tools. Inconsistent results were found in studies using standardised tools, placing doubt of the reliability of these tools in the nursing context. On examination of the domain concepts construct validity was questionable with several non-standardised tools used.

Nursing and midwifery education needs to prepare graduates to work effectively in complex, fast paced and uncertain environments. Continued collection of data using measures of

generalised CT is unlikely to help improve curricula, teaching methods, or preparation of students for professional practice. There is a need to develop discipline specific instruments to measure CT in nursing and midwifery, and more specifically tools that measure the application of CT to practice. Considering the complexity of critical thinking in nursing and midwifery practice, and that CT development occurs over a long time, measurement requires a long term, multi-method approach over this time.

References

Anderson, N., Booth, L., Catalano, J., Gaines, L., Horner, M., % McCormick, S., (2000). *Critical thinking process test: Development and technical eport*. Stillwell, KS: Educational Resource, Inc.

Assessment Technologies Institute, LLC. (2001). CT assessment: Developmental and statistical report. Overland Park, KS: Author.

Abel, W. M., & Freeze, M. (2006). Evaluation of concept mapping in an associate degree nursing program. *Journal of Nursing Education, 45*(9), 356-364.

Atay, S., & Karabacak, Ü. (2012). Care plans using concept maps and their effects on the critical thinking dispositions of nursing students. *International Journal of Nursing Practice, 18*(3), 233-239. doi: 10.1111/j.1440-172X.2012.02034.x

Beckie, T. M., Lowry, L. W., & Barnett, S. (2001). Assessing critical thinking in baccalaureate nursing students: a longitudinal study. *Holistic Nursing Practice, 15*(3), 18-26.

Brown, J. M., Alverson, E. M., & Pepa, C. A. (2001). The influence of a baccalaureate program on traditional, RN-BSN, and accelerated students' critical thinking abilities. *Holistic Nursing Practice, 15*(3), 4-8.

Brunt, B. A. (2005). Models, measurement, and strategies in developing critical-thinking skills. *Journal of Continuing Education in Nursing, 36*(6), 255-262.

CASP (2013). *Critical Thinking Appraisal Skills Programme:CASP Checklists.* UK: CASP. Retrieved from http://www.casp-uk.net/#!casp-tools-checklists/c18f8

Chau, J. P. C., Chang, A. M., Lee, I. F. K., Ip, W. Y., Lee, D. T. F., & Wootton, Y. (2001). Effects of using vidoetaped vignettes on enhancing students' critical thinking ability in a baccalaureate nursing programme. *Journal of Advanced Nursing, 36*(1), 112-119. doi: 10.1046/j.1365-2648.2001.01948.x

Chen, F., & Lin, M. (2003). Effects of a nursing literature reading course on promoting critical thinking in two-year nursing program students. *Journal of Nursing Research (Taiwan Nurses Association), 11*(2), 137-147.

Cheng, Y. Y., Wang, W. C., Wu, J. J., & Hwang, C. K. (1996). A preliminary report on theconstruction of the critical thinking scale (in Chinese). *Psychological Testing,* 43, 213-226.

Choi, E., Lindquist, R., & Song, Y. (2014). Effects of problem-based learning vs. traditional lecture on Korean nursing students' critical thinking, problem-solving, and self-directed learning. *Nurse Education Today, 34*(1), 52-56. doi: 10.1016/j.nedt.2013.02.012.

Chou, F.H., Jian, S.Y., Tseng, H.C., Ko, H.G., (2004). *The evaluation of students performance in applying problem-based learning to a nursing course.* The Research Outcome of National Science Council, Taiwan.

Daley, B.J., Shaw, C.R., Balistrieri, T., Glasenapp, K., & Piacentine, L. (1999). Concept maps: A strategy to teach and evaluate critical thinking. *Journal of Nursing Education*, 38, 42-47.

Daly, W. M. (2001). The development of an alternative method in the assessment of critical thinking as an outcome of nursing education. *Journal of Advanced Nursing, 36*(1), 120-130. doi: 10.1046/j.1365-2648.2001.01949.x

Dehkordi, A. H., & Heydarnejad, M. S. (2008). The effects of problem-based learning and lecturing on the development of Iranian nursing students' critical thinking. *Pakistan Journal of Medical Sciences, 24*(5), 740-743.

DeSimone, B. B. (2006). Curriculum design to promote the critical thinking of accelerated bachelor's degree nursing students. *Nurse Educator, 31*(5), 213-217.

Evans, B. C., & Bendel, R. (2004). Cognitive and ethical maturity in baccalaureate nursing students: did a class using narrative pedagogy make a difference? *Nursing Education Perspectives, 25*(4), 188-195.

Facione, P. A. (1990). *Critical Thinking: A Statement of Expert Consensus for Purposes of Educational Assessment and Instruction, Executive Summary: "The Delphi Report".*

CA: The Californian Academic Press. Retrieved from http://assessment.aas.duke.edu/documents/Delphi_Report.pdf
Facione, P. A. & Facione, N. C. (1992a). *The California Critical Thinking Dispositions Inventory (CCTDI); and the CCTDI Test manual.* Millbrae, CA: California Academic Press.
Facione, P. A., & Facione N. C. (1992b). *The California Critical Thinking Skills Test: Test Manual.* Millbrae, CA: California Academic Press.
Facione, P., & Facione, N. (1994). *The California Critical Thinking Disposition Inventory (CCTDI): Test Manual*. Millbrae, CA: California Academic Press.
Facione, N. C, & Facione, P. A. (1996). Assessment design issues for evaluating critical thinking in nursing. *Holistic Nursing Practice,* 10(3), 41-53.
Facione, P. A., & Facione, N. C. (1998). *The California Critical Thinking Skills Test: CCTST test manual.* Millbrae, CA: California Academic Press.
Facione, P. A., Facione, N. C., & Winterhalter, K. (2010). *The Health Sciences Reasoning Test: Test manual.* Millbrae, CA: California Academic Press.
Fountain, L. (2011). *Thinking Like a 21st Century Nurse: Theory, Instruments, and Methodologies for Measuring Clinical Thinking.* Paper presented at the Annual Meeting of the American Educational Research Association New Orleans, University of Maryland.
Goodstone, L., Goodstone, M. S., Cino, K., Glaser, C. A., Kupferman, K., & Dember-Neal, T. (2013). Effect of simulation on the development of critical thinking in associate degree nursing students. *Nursing Education Perspectives, 34*(3), 159-162.
Insight Assessment (2013). California Critical Thinking Disposition Inventory (CCTDI). San Jose, CA.: The Californian Academic Press Retrieved from: http://www.insightassessment.com/Products/Products-Summary/Critical-Thinking-Attributes-Tests/California-Critical-Thinking-Disposition-Inventory-CCTDI#sthash.HBvLVC1c.dpbs
Jones, M. (2008). Developing clinically savvy nursing students: an evaluation of problem-based learning in an associate degree program. *Nursing Education Perspectives, 29*(5), 278-283.
Jun, W. H., Lee, E. J., Park, H. J., Chang, A. K., & Kim, M. J. (2013). Use of the 5E learning cycle model combined with problem-based learning for a fundamentals of nursing course. *Journal of Nursing Education, 52*(12), 681-689.
Kakai H. (2003). Re-examining the factor structure of the California Critical Thinking Disposition Inventory. *Perceptual and Motor Skills*, 96(2), 435-438.
Kaplan, R. M., Sacuzzo, D.P. (1997) *Psychological Testing: Principles, Applications and Issues* (4th ed.), Pacific Grove, CA: Brooks/Cole.
Lawson, S. B. (2012). The Effectiveness of Concept Mapping as an Educational Tool to Enhance Critical Thinking Skills in Undergraduate Nursing Students. Unpublished Thesis. Indiana: Ball State University.
L'Eplattenier, N. (2001). Tracing the development of critical thinking in baccalaureate nursing students. *Journal of the New York State Nurses Association, 32*(2), 27-32.
Lee, W., Chiang, C. H., Liao, I. C., Lee, M. L., Chen, S.L., & Liang, T. (2013). The longitudinal effect of concept map teaching on critical thinking of nursing students. *Nurse Education Today, 33*(10), 1219-1223. doi: 10.1016/j.nedt.2012.06.010
Mann, J. (2012). Critical Thinking and Clinical Judgment Skill Development in Baccalaureate Nursing Students. *Kansas Nurse, 87*(1), 26-31.
Mong-Chue, C. (2000). Professional issues. The challenges of midwifery practice for critical thinking. *British Journal of Midwifery, 8*(3), 179-183.
Morey, D. J. (2012). Development and Evaluation of Web-Based Animated Pedagogical Agents for Facilitating Critical Thinking in Nursing. *Nursing Education Perspectives, 33*(2), 116-120. doi: 10.5480/1536-5026-33.2.116
Muoni, T. (2012). Decision-making, intuition, and the midwife: Understanding heuristics. *British Journal of Midwifery, 20*(1), 52-56.

Naber, J., & Wyatt, T. H. (2014). The effect of reflective writing interventions on the critical thinking skills and dispositions of baccalaureate nursing students. *Nurse Education Today, 34*(1), 67-72. doi: 10.1016/j.nedt.2013.04.002

Park, S.H., (1999). The effects of the program for the improvement of college students' critical thinking ability [Korean]. *Journal of Educational Psychology,* 13 (4), 93–112.

Park, J.A., & Kim, B.J. (2009). Critical thinking disposition and clinical competence in general hospital nurses [Korean]. *Journal of Korean Academy of Nursing*, 39, 840-850.

Paul, R. W. (1993). *Critical Thinking.* Santa Rosa, CA; Foundationfor Critical Thinking.

Pucer, P., Trobec, I., & Žvanut, B. (2014). An information communication technology based approach for the acquisition of critical thinking skills. *Nurse Education Today, 34*(6), 964-970. doi: 10.1016/j.nedt.2014.01.011

Ravert, P. (2008). Patient simulator sessions and critical thinking. *Journal of Nursing Education, 47*(12), 557-562. doi: 10.3928/01484834-20081201-06

Rubenfeld, M. G. & Scheffer,B. K. (2006). *Critical thinking TACTICS for nurses : tracking, assessing, and cultivating thinking to improve competency-based strategies***.** Sudbury, Mass : Jones and Bartlett.

Scheffer, B. K., & Rubenfeld, M. G. (2000). A consensus statement on critical thinking in nursing. *Journal of Nursing Education, 39*(8), 352-359.

Scholes, J., Endacott, R., Biro, M., Bulle, B., Cooper, S., Miles, M., Gilmour, C., Buykx, P., Kinsman, L., Boland, R., Jones, J., Zaidi, F. 2012. Clinical decision-making: midwifery students' recognition of, and response to, post partum haemorrhage in the simulation environment. BMC Pregnancy and Childbirth 12, 19.

Seldomridge, L. A., & Walsh, C. M. (2006). Measuring critical thinking in graduate education: what do we know? *Nurse Educator, 31*(3), 132-137.

Shin, K. R., Lee, J. H., Ha, J. Y., & Kim, K. H. (2006). Critical thinking dispositions in baccalaureate nursing students. *Journal of Advanced Nursing, 56*(2), 182-189. doi: 10.1111/j.1365-2648.2006.03995.x

Shinnick, M. A., & Woo, M. A. (2013). The effect of human patient simulation on critical thinking and its predictors in prelicensure nursing students. *Nurse Education Today, 33*(9), 1062-1067. doi: 10.1016/j.nedt.2012.04.004

Spelic, S. S., Parsons, M., Hercinger, M., Andrews, A., Parks, J., & Norris, J. (2001). Evaluation of critical thinking outcomes of a BSN program. *Holistic Nursing Practice, 15*(3), 27-34.

Stevens, J., Brenner, C., & Brenner, Z. R. (2009). The peer active learning approach for clinical education: a pilot study. *Journal of Theory Construction & Testing, 13*(2), 51-56.

Stewart, S., & Dempsey, L. F. (2005). A longitudinal study of baccalaureate nursing students' critical thinking dispositions. *Journal of Nursing Education, 44*(2), 81-84.

Stone, C. A., Davidson, L. J., Evans, J. L., & Hansen, M. A. (2001). Validity evidence for using a general critical thinking test to measure nursing students' critical thinking. *Holistic Nursing Practice, 15*(4), 65-74.

Sullivan-Mann, J., Perron, C. A., & Fellner, A. N. (2009). The effects of simulation on nursing students' critical thinking scores: a quantitative study. *Newborn and Infant Nursing Reviews, 9*(2), 111-116.

Tiwari, A., Lai, P., So, M., & Yuen, K. (2006). A comparison of the effects of problem-based learning and lecturing on the development of students' critical thinking. *Medical Education, 40*(6), 547-554.

Tseng, H. C., Chou, F. H., Wang, H.H., Ko, H.K., Jian, S.Y., & Weng, W.C. (2011). The effectiveness of problem-based learning and concept mapping among Taiwanese registered nursing students. *Nurse Education Today, 31*(8), e41-46. doi: 10.1016/j.nedt.2010.11.020

Walsh, C. M, & Hardy, R. C. (1997) Factor structure stability of the California Critical Thinking Disposition Inventory across gender and various student majors. *Perceptual and Motor Skills,* 85,1211-1228.

Walsh, C. M., & Seldomridge, L. A. (2006). Measuring critical thinking: one step forward, one step back. *Nurse Educator, 31*(4), 159-162.
Watson, G., & Glaser, E. M. (1980). *Watson-Glaser Critical Thinking Appraisal*. San Antonio, TX: Psychological Corp.
Wheeler, L. A., & Collins, S. K. R. (2003). The influence of concept mapping on critical thinking in baccalaureate nursing students. *Journal of Professional Nursing, 19*(6), 339-346.
Wood, R. Y., & Toronto, C. E. (2012). Measuring Critical Thinking Dispositions of Novice Nursing Students Using Human Patient Simulators. *Journal of Nursing Education, 51*(6), 349-352. doi: 10.3928/01484834-20120427-05
Yeh, M., & Chen, H. (2005). Effects of an educational program with interactive videodisc systems in improving critical thinking dispositions for RN-BSN students in Taiwan. *International Journal of Nursing Studies, 42*(3), 333-340.
Yu, D., Zhang, Y., Xu, Y., Wu, J., & Wang, C. (2012). Improvement in critical thinking dispositions of undergraduate nursing students through problem-based learning: a crossover-experimental study. *Journal of Nursing Education, 52*(10), 574-581.
Yuan, H., Kunaviktikul, W., Klunklin, A., & Williams, B. A. (2008). Improvement of nursing students' critical thinking skills through problem-based learning in the People's Republic of China: a quasi-experimental study. *Nursing & Health Sciences, 10*(1), 70-76.
Zadeh, H. H., Khajeali, N., Khalkhali, H., & Mohammadpour, Y. (2014). Effect of evidence-based nursing on critical thinking disposition among nursing students. *Life Science Journal, 11* (9 Spec. Issue), 487-491.