CLEAR Exam Review

Volume XVIII, Number 1
Winter 2007

A Journal

CLEAR Exam Review

Volume XVIII, Number 1 Winter 2007

CLEAR Exam Review is a journal, published twice a year, reviewing issues affecting testing and credentialing. CER is published by the Council on Licensure, Enforcement, and Regulation, 403 Marquis Ave., Suite 100, Lexington, KY 40502.

Editing and composition of this journal have been underwritten by Prometric, which specializes in the design, development, and full-service operation of high-quality licensing, certification and other adult examination programs.

Subscriptions to CER are sent free of charge to all CLEAR members and are available for $30 per year to others. Contact Stephanie Thompson at (859) 269-1802, or at her e-mail address, sthompson@mis.net, for membership and subscription information.

Advertisements and Classifieds (e.g., position vacancies) for CER may be reserved by contacting Stephanie Thompson at the address or phone number noted above. Ads are limited in size to 1/4 or 1/2 page, and cost $100 or $200, respectively, per issue.

Editorial Board: Janet Ciuccio, American Psychological Association; Steven Nettles, Applied Measurement Professionals; Jim Zukowski, Texas State Board of Dental Examiners.

Coeditor
Michael Rosenfeld, Ph.D.
Educational Testing Service
Princeton, New Jersey 08541-0001
mrosenfeld@ets.org

Coeditor
F. Jay Breyer, Ph.D.
Prometric
2000 Lenox Drive, Lawrenceville, NJ 08648
jay.breyer@prometric.com

Contents

Copyright © 2007 Council on Licensure, Enforcement, and Regulation. All rights reserved.
ISSN 1076-8025

Assessing Critical Thinking Using a Talk-Aloud Protocol

Anne Wendt, Ph.D., RN, CAE
Lorraine E. Kenny, MS, RN
Casey Marks, Ph.D.

"With the importance placed on the ability of the entry-level practitioners to think critically, regulators may want to consider the use of alternate format items in addition to multiple-choice items to assess higher order thinking skills."

Abstract
This article presents the results of a research study that addresses how entry-level nurses process information in order to answer multiple-choice items versus a paired item that uses an alternate format. A talk-aloud protocol was used to assist in the identification of the cognitive processing that was required by entry-level nurses as they responded to items. A purposeful sample was used to select the participants: seven registered nurses and five practical nurses licensed for less than a year. Results of the study suggest that some items that use alternate formats require participants to use higher cognitive processing than a paired multiple-choice item. With the importance placed on the ability of the entry-level practitioners to think critically, regulators may want to consider the use of alternate format items in addition to multiple-choice items to assess higher order thinking skills.

Effective clinical decision-making is one of the most important contributions made by health care professionals (White, 2003). A large component of effective clinical decision-making, and thus of the successful practitioner, centers on the ability to understand complex issues and to think critically. Using items from the National Council Licensure Examinations for Registered and Practical Nurses (NCLEX-RN® and NCLEX-PN®), a study was undertaken to determine whether different levels of cognitive processing, such as the higher order skills of critical thinking and clinical decision-making, were used by entry-level nurses to answer examination items of varying formats and content areas. A qualitative method using a talk-aloud protocol was used to investigate how entry-level nurses process information in order to answer a question posed to them. The results of this study may provide insight into ways regulators can assess critical thinking and clinical decision-making as part of their licensure process.

Cognitive Processing

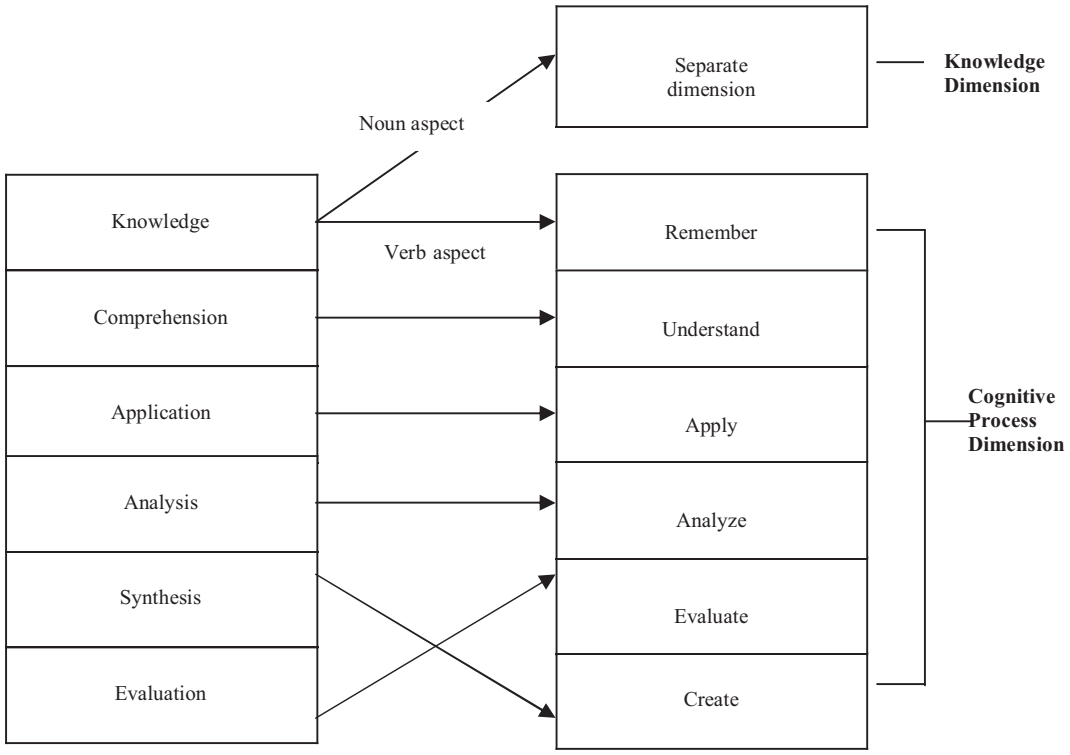
Many methods of evaluating an individual's skills and abilities in a domain of knowledge involve evaluating that individual's cognition—often referred to as cognitive processing ability. Various taxonomies have been developed in an attempt to categorize the different levels of cognitive processing that are used to answer test items. Probably the most well known taxonomy used to categorize educational objectives and then test items is Bloom's taxonomy (Bloom, 1956).

Bloom's taxonomy contains six major classes from lowest level of cognitive processing to highest level of cognitive processing: Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation (Bloom, 1956). Bloom's taxonomy is a cumulative hierarchy. This taxonomy assumes that the more complex cognitive behaviors include the simpler cognitive behaviors. Items categorized at successively higher levels of the taxonomy require more cognitive processing to answer a question. Several research studies provide evidence that supports this cumulative hierarchy and thus the ordering of the less complex categories of Knowledge/Recall, Comprehension, and Application. (Anderson & Krathwohl, 2001; Miller, et al., 1991; Buckwalter, et al., 1981). However, when the higher levels of cognitive processing are considered, there seems to be weaker empirical evidence for the hierarchical model (Anderson & Krathwhol, 2001). Indeed, some well-known authors have

Anne Wendt, Lorraine E. Kenny, and Casey Marks are on the staff of the National Council of State Boards of Nursing, Inc.

revised Bloom's taxonomy as can be seen in Figure 1. (Anderson, et al., 2001, p. 310)

Figure 1. Comparison of Bloom's Framework to Revised Framework



In exploring this revision to the taxonomy, cognitive complexity appeared to be the ordering principle for Bloom's taxonomy and Create, which uses inductive reasoning, is inherently a more complex process than Evaluate, which uses deductive reasoning. In Create, the examinee gathers information and views it in light of personal knowledge and experience. The cognitive process used in Create involves putting elements together to form a coherent or functional whole by reorganizing elements into a new pattern or structure. In using deductive reasoning for the Evaluate category, the examinee provides assertions that predictably lead to a conclusion measuring soundness; makes judgments based on criteria and standards.

In general there seems to be evidence to support the use of Bloom's taxonomy for the "lower" levels of cognitive processing and most especially for use in categorizing closed response items such as multiple-choice items. However, when constructed responses of some of the alternate format items are considered, there may be some challenges. In fact there may be challenges to any taxonomy that is used to categorize items (Osterlind & Merz, 1994). In this study the revised taxonomy is used when categorizing the cognitive processing used to answer items.

Critical Thinking

Developing the ability to think critically is essential to nursing practice as it is with any profession where there is an obligation to be licensed to practice. The Delphi Research Project of 1990 describes the attributes of an ideal critical thinker, and authors (Facione, Facione & Sanchez, 1994) also contend these are the characteristics of a nurse with ideal clinical judgment using the core critical thinking cognitive skills of interpretation, analysis, inference, evaluation and explanation as an interactive, reflective, reasoning process. Although not easy to characterize, critical thinking is often thought of as a collection of mental skills that can

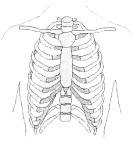
be taught and therefore assessed within the confines of test development. Evaluating and predicting are two aspects of critical thinking for which we can prepare test items (Haladyna, 1997).

Many test developers propose that multiple-choice items can be used to evaluate critical thinking, as long as the items are focused on measuring higher-order thinking ability. McDonald (2002) proposes that such an assessment consists of the ability to use item information in a unique situation—moving away from recall or comprehension-level questions that require only rote memorization skills. This unique situation is considered to be enhancement of a recall question into an application or analysis type of question. While there is agreement that item development at the application and analysis level is fundamental to the measurement of critical thinking, there appears to be increased evidence of the use of higher-order thinking when examinees answer constructed-response items. Items not limited to a single correct answer encourage the examinee to move from recall to application/analysis and therefore demonstrate cognitive processes that can be identified as critical thinking (McDonald, 2002).

Alternate Item Types

In 1994 the NCLEX® examinations moved from a paper and pencil format of multiple-choice items to computer-adaptive technology using those same item formats. Innovations in computer-based testing include additional item types with features that include sound, graphics, animation and video integrated into the item stem, response options or both. The computer interface for items has moved from multiple-choice type items of selecting one answer from several response alternatives, to the ability to drag and drop objects in order to rank answer options, click on graphics, and choose multiple-correct responses. In addition, advances have been made in the scoring of fill-in-the-blank items and essays. The following pages describe the types of alternate item formats used in this study. These items were paired with multiple-choice items to determine the ability of alternate items to tap into higher-order thinking. Figure 2 contains a comparison of traditional item types and alternate item formats. A discussion of the various item types follows.

Figure 2. Comparison of Traditional and Alternate Item Formats

Traditional Item Type	Alternate Item Type
Multiple Choice Item: While assessing the patient's abdomen, which of the following sequences of the examination should the nurse recognize as appropriate? 1. inspection, auscultation, percussion & palpation. 2. auscultation, palpation, inspection, & percussion 3. inspection, palpation, auscultation, & percussion. 4. palpation, inspection, auscultation, & percussion	**Fill in the Blank/Ordered Response:** While assessing the patient's abdomen, in what sequence should the examination be conducted (identify steps by inserting the number of the first steps, second step, etc.)? ___ test for rebound tenderness ___ percussion ___ auscultation ___ palpation ___ inspection
Multiple Choice Item: When caring for a client who has a wound infected with methicillin-resistant Staphylococcus aureus (MRSA), which of the following infection control procedures should the nurse implement? 1. Place the client in a private room. 2. Ask the client to wear a surgical mask. 3. Use sterile gloves to remove the wound dressing. 4. Wear a protective gown when providing wound care.	**Multiple Response Item:** When caring for a client who has a wound infected with methicillin-resistant Staphylococcus aureus (MRSA), which of the following infection control procedures should the nurse implement? (Check all that apply) ___ 1. Wear a protective gown when entering the client's room ___ 2. Put on a particulate respirator mask when administering medications to the patient ___ 3. Wear gloves when delivering the clients meal tray. ___ 4 Ask the client's visitors to wear a surgical mask when in the client's room ___ 5. Wear sterile gloves when removing the client's dressing. ___ 6. Put on a face shield before irrigating the client's wound.
Multiple Choice Item: The nurse is performing a cardiac assessment upon admission. Which of the following describes the best anatomic location to auscultate the mitral valve at its loudest? 1. Second intercostal space at the right sternal border. 2. Third intercostal space at the left mid-clavicular line. 3. Fourth intercostal space at the left sternal border. 4. Fifth intercostals space at the left mid-clavicular line.	**Hot Spot:** The nurse is performing a cardiac assessment upon admission. Click on the area where the nurse should auscultate to hear the mitral valve at its loudest? 
Multiple Choice Item: The nurse is caring for a client whose intake and output must be calculated. The nurse observes that the client has consumed 8 ounces of apple juice, one hamburger on a bun, one-half cup of green beans, 8 ounces of tea, and one cup of ice cream. Which of the following should the nurse record as the client's intake? 1. 360 milliliters 2. 560 milliliters 3. 720 milliliters 4. 760 milliliters	**Fill in the Blank/Calculation:** The nurse is monitoring the dietary intake and output of a client. The nurse observes that the client has consumed 8 ounces of apple juice, one hamburger on a bun, one-half cup of green beans, 8 ounces of tea, and one cup of ice cream. How many milliliters should the nurse record for the client's intake? _____ milliliters
* Additional item types using audio and multi-screen chart formats were included but are not part of this example.	

Fill-in-the-Blank (FIB) items are examples of constructed responses where, unlike the selected response of standard multiple-choice items, the examinee is not given a list of responses to choose from as the correct answer. An example of this type of item within the research examination is the fill-in-the-blank 'calculation' item (FBC). Nursing proficiency in calculation is a vital aspect of medication administration including calculation of medication doses for oral and parenteral administration. In addition, nurses need to know how to calculate intake/output as part of body fluid volume.

Using this item format presents some challenges. Specific scoring rules must be developed prior to testing, allowing each item to have a list or a range of correct answers. In the area of medication content, the acceptable answers must remain within the realm of safe nursing/medical practice, rather than strict arithmetic calculations. Rounding techniques, although essentially correct, could impact patient safety and/or prescriptive instructions if it gives the client too little or too much medication.

Another short-answer item type used within this test is the Ordered Response item. In this item type an examinee ranks a set of response options in the correct order. Examinees are presented with a list of essential steps to a nursing procedure (e.g., cardiopulmonary resuscitation-CPR) and asked to rank order the steps in the correct sequence. Upon deciding the correct sequence, the examinee lists the numbers in the correct order in the answer box.

Items using graphics have been a part of the NCLEX® examination since the days of paper-and-pencil testing and continue to function within the current computer environment. The use of graphics is considered one of the most common non-text media used in computerized testing. The graphics themselves can be used as all or part of individual items—either in the question itself or as part of the response option. A Hot Spot Item depicts the area on the graphic that best answers the question posed.

While traditional multiple-choice items allow the examinee to select a response from a list of four options, the Multiple-Response (MR) alternate item is a variant on this item type that allows the examinee to choose 'all that apply'. These variant models are used without cueing the examinee to the actual number of correct responses. Additionally these MR formats do require that the examinee have the ability to discriminate from a list of important content implications (Jodoin, 2003). Within nursing content, this item type identifies the examinees ability to consider all possibilities in providing patient care in a given situation. Depending on the phrasing of the content in the item, the nurse may be required to discriminate between non-mutually exclusive actions that would impact the outcome of patient care.

One of the competencies required by the registered nurse is the ability to perform physical assessments of the client. Without an actual psychomotor skills component to the nursing licensure exam, current multiple-choice items are unable to assess a candidate's competence in identifying such things as lung sounds, heart sounds or the non-verbal cues communicated in nurse-client interactions. By including audio and video clips or pictures to items, it may be possible to assess candidate competence in these areas. In addition, it may be possible to decrease the candidate reading load during an examination. Acknowledging this flexibility, Parshall, et al., (2000) state: "…a major advantage of administering tests via computer is the opportunity to include non-text media in the items. The use of these media can reduce dependence on reading skills, as well as enhance the validity and task congruence of a test." (p. 136) The development of assessments of listening skills might also be important because the visual and audio channels of communication tap different cognitive processes. For example, there is evidence that while audio information places greater demand on short-term memory, multiple streams of information can be processed concurrently more easily and accurately when communicated aurally (Fitch and Kramer, 1994 in Parshall, et al.).

For the purpose of this research, audio items were simulated using a wave file. The examinee, using earphones to amplify sound quality, could then 'listen to a sound'. The sound could be repeated as often as necessary by clicking on the audio player on the computer.

The use of clinical scenarios in items provides the opportunity for more authentic depictions of patient situations. The creation of clinical scenarios for patient situations can consist of high volume, high risk, problem prone situations, and areas where the nurse is asked to apply concepts and theories. For the alternate items created for this research, clinical scenarios were created and the examinee was asked to select an answer from a long list of options. In the clinical scenario the examinee would review various history and physical notations in the patient chart, as well as current laboratory and clinical data. To review the chart, selection of a tab denoting a section of the chart would enable a window to appear with the concomitant information. The examinee was able to move back and forth in the sections of the chart and then determine the most correct answer. While the format for answering the questions was actually selected response, the context-dependent items provided the ability to assess the examinee's understanding of which actions to take in a clinical situation based on the knowledge of system functioning as determined in the information provided.

As has been discussed, there is a great deal of information on how individuals process information and learn. In addition, there are many taxonomies that can be used to categorize cognitive processing and with the categorization of cognitive processing, the manifestation of that processing for assessment purposes. In nursing and other professions, critical thinking has been identified as an important trait to develop. Measurement of that trait is difficult but is valued by educators and regulators. One method that has been identified as useful in identifying critical thinking is the use of alternate items. However, it remains uncertain as to whether it is possible to objectively measure critical thinking, and if the items that employ alternate formats actually require higher cognitive processing when compared to a paired multiple-choice item. This study is intended to determine if alternate items are assessing higher levels of cognitive processing.

Methodology

The design of this study uses a non-experimental research method involving a talk-aloud protocol to identify the cognitive processing required to answer alternate item formats paired to multiple-choice formats. Based on the work of Ericsson and Simon (1984) the assumption is that it is possible for subjects to report on their last cognitive processes based on information and cues retrieved from their short term memory. Talk-aloud protocols use the subject's verbal reports as data in determining the cognitive processes, with the addition of retrospective reporting to verify the data. Experts then evaluate the talk alouds and assign a cognitive level to the verbal report. The results of the experts rating the talk alouds are then analyzed using a multi-faceted variation of the Rasch measurement model, FACETS (Linacre, 2003). FACETS can be used to jointly measure people, items and raters. FACETS produces "measures" of persons and items. The "measures" of the participants' cognitive processing of items as defined by the experts' ratings are then used when comparing the item pairs.

In order to identify whether there was a difference between the cognitive processes of alternate items as compared to conventional multiple-choice items, it was necessary to employ the use of verbal reports. The 'think-aloud' theory devised by Newell & Simon (1972, in Taylor, 2000) proposes that it is possible to record and to identify the problem-solving strategies in use during exercises. Using this protocol, in which the examinee is asked to relate what is in his/her mind while working through the item, provides the ability to identify the actual cognitive processing taking place.

For the purpose of various talk-aloud research, text types may include reading paragraph comprehension, where easy well-written texts may not provide suitable verbalization—but rather a reproduction of the text itself. When text becomes more difficult due to unfamiliar topics, poor organization or unfamiliar writing styles, talk aloud can produce more information than only the reproduction of text (Katalin, 2002). In this research reading has a specific purpose and the test question formats described earlier make the activity of reading and thinking in order to find the correct response suitable for talk-aloud research.

Talk Aloud Study

Nurses within a thirty-mile area of our offices (in Chicago, IL) who had successfully passed the NCLEX® examination and were within their first year of practice were asked to participate in this study. Participants signed a confidentiality agreement and were given a tutorial on the talk-aloud protocol methodology, asking them to talk-aloud constantly from the time a problem was presented until they had given their final answer. Finally, the participants were asked about their educational background, how many times they had taken the NCLEX® examination before passing, and their computer skills to assure a mix of skill levels.

They were instructed not to plan out what they were to say or try to explain their reasoning to the researcher. The participants were also told that if they were silent for a long period of time the researcher would remind them to talk. A series of warm-up exercises were presented including simple multiplication problems and anagrams until participants were comfortable with the talk-aloud procedures. For practice in retrospective reporting, the study participants were asked to talk aloud and identify the number of windows in their parent's house. Next they were asked to describe how they were thinking as they arrived at that number. The same type of practice was used to identify twenty animals. The practice sessions continued until the participants were comfortable with the protocol.

The method required that the verbalizations were recorded and the researchers did not interfere with the process. The tape recorder was placed to one side and the researcher sat behind and out of sight of the examinee, who was working at one laptop with the examination and with the audio recordings. The participant was given a tutorial on the alternate item types with samples of each question type and a full explanation of the specific operation of the computer being used. Throughout the recording session the researcher would prompt the subjects to "keep talking" whenever there was an extended period of silence.

Following completion of the research test, the participants were given a copy of the items just completed and asked to retrospectively state what they were thinking about when

they answered the item. They were instructed not to re-answer the question, but to review what they were thinking about when they originally answered. Ericsson and Simon, (1984) state that "even for cognitive processes of long duration, where we know that the retrospective report will be incomplete, it will be quite useful…it will more clearly convey the general structure of the process…" (p. 379)

Once completed, the recordings were transcribed verbatim. Three experts evaluated the talk aloud transcription and categorized the findings according to the six cognitive levels: knowledge, comprehension, application, analysis, evaluate and create. The decision to stop the talk alouds was based upon the determination by the experts as to whether or not new information was being obtained from the review. The experts were to evaluate the talk alouds of 5-10 entry level RNs and 5-10 entry level LPNs. Should there be no new information gained from the talk alouds after 5 cases, the talk alouds and transcriptions could be stopped. After analysis of the transcription of five LPNs and seven RNs it was determined that no new information would be gained.

Expert Rating

Three nursing experts were used to evaluate the cognitive processing of the participants talk aloud as they answered the items. The raters received information on the purpose of the study, training materials addressing talk-aloud protocol and cognitive processing, and background material. All of the raters had been involved with the NCSBN Examination Committee and were very familiar with cognitive processing and NCLEX® items. Prior to evaluating the transcriptions, the raters were oriented and a sample transcription was evaluated independently, and then discussed as a group to bring forward any issues regarding the scoring matrix that was developed. The experts were asked to evaluate the transcript independently, using the scoring matrix. They were asked to identify lines from the transcript that supported their rating of the cognitive processing of the items in the research test. The experts were "blinded" as to the sequencing of the items in the research test in order to prevent bias in their ratings of cognitive processing.

In developing the research test, care was taken to vary not only the item formats, but within formats to vary the nursing content as well as expected levels of item difficulty. In addition, to manage potential cueing and impact of cueing on cognitive processing levels, the positions of the alternate items and the multiple-choice items in the test were juxtaposed. All items used in the study were validated in nursing textbooks required by entry level nursing programs and were evaluated by item review panels who certified the correct answer and that the items are appropriate for entry level practitioners. This information can be used to determine if the different nursing content areas and competencies being assessed impacted the item format and/or cognitive processing of the items. Thus, a variety of content areas including several items on the same content were used for the various item types. It was expected that the study would be able to determine that cognitive processing for item formats is not necessarily related to content of item. In working with calculation items, participants were provided with calculators just as they are in the NCLEX® examinations.

As discussed, a purposeful sample was used to select the participants for the Talk-Aloud research. Seven registered nurses and five practical nurses participated in the Talk Aloud. All of the Talk Aloud participants were female and were in practice for less than one year. The ethnicity of the participants included White non-Hispanic, Hispanic, Black, and Asian.

The rating scale used in this study is six categories of cognitive processing skills:1=Recall; 2=Comprehend; 3=Apply; 4=Analyze; 5=Evaluate; 6=Create. Three expert judges read the talk aloud transcripts of each candidate and rated the cognitive process used to answer each test item. Objective measurement assumes each rater's individuality and is not concerned with inter-rater reliability as an end to itself--rater severity is only one of many indicators. Rather, it is the consistency with which each judge uses the evaluation form that is important. A FACETS analysis will adjust for the different types and severity of raters as long as they share a common understanding of the evaluation form and are individually consistent in their use of the rating scale (Linacre, 2003).

Results

As indicated by an analysis of the raters, items, and talk aloud participants using FACETS, the RN raters have similar views when rating the talk-aloud transcripts as revealed by the low separation of .21 logits between the raters. PN raters are more variable with a range of nearly a logit difference in how they rate the talk aloud participants. Fit statistics show all the raters are internally consistent in their judgments of cognitive processing skills as described by the candidates. Using a more conventional index of rater reliability, the rater agreement is 58% for PN and 46% for RNs, which is quite good for this small sample.

For RNs and PNs, the lower levels of cognitive processing, recall, comprehend and apply account for two thirds of the ratings; apply and analysis for one third; with less than one percent using evaluate; and no one create, the highest cognitive processing levels. Since half of the items are multiple-choice items, which tend to measure lower levels of cognitive processing, this is not an unexpected finding. When alternate items versus paired multiple-choice items are

examined, a very different picture emerges.

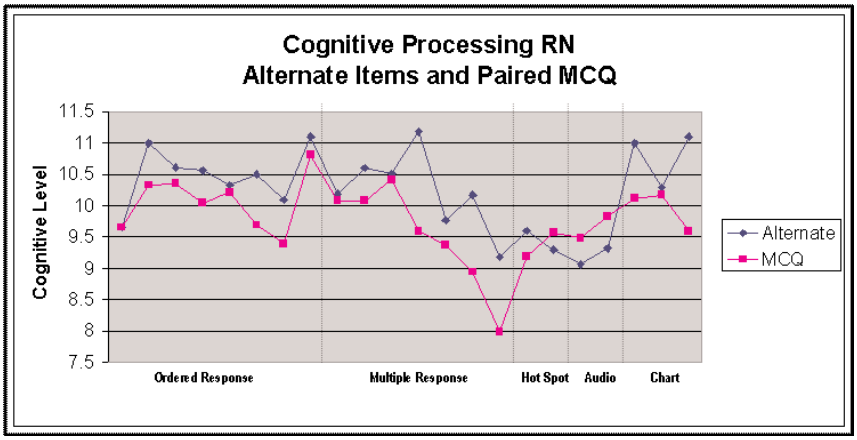
Table 1: Comparison of RN Alternate and Multiple Choice Items

RN							
Alternate Items				**Multiple-Choice Items**			
	Measure	**SE**	**Label**		**Measure**	**SE**	**Label**
Ordered Response				**Ordered Response**			
1	9.65	0.29	Ab Abscess	1	9.65	0.29	ab abscess
2	11	0.22	CPR	2	10.33	0.24	CPR prioritize
14	10.61	0.23	CPR infant	14	10.35	0.24	CPR infant
17	10.56	0.23	trach care	17	10.04	0.26	trach care
21	10.33	0.24	VS/labs-action	21	10.21	0.25	VS/labs-action
22	10.5	0.24	tube feeding	22	9.69	0.3	tube feeding
Multiple Response				**Multiple Response**			
3	10.09	0.25	OB-graphic-VarDec	3	9.39	0.31	OB-VarDec
5	11.1	0.22	assignments to LPN	5	10.81	0.22	assignments to LPN
8	10.19	0.24	infection control	8	10.08	0.25	infections control
10	10.6	0.23	late decels/labor	10	10.08	0.25	late decels/labor
12	10.51	0.22	assignments	12	10.41	0.23	assignments
16	11.18	0.23	heart failure	16	9.59	0.3	breath sounds
18	9.76	0.28	newborn assess	18	9.37	0.33	newborn assess
20	10.17	0.25	pressure ulcer	20	8.94	0.36	pressure ulcer
Hot Spot				**Hot Spot**			
7	9.18	0.33	aortic valve	7	7.99	0.38	aortic valve
11	9.6	0.31	chest drainage system	11	9.19	0.34	chest drainage system
Audio				**Audio**			
4	9.29	0.32	crackles	4	9.57	0.3	crackles
9	9.06	0.35	vesicular	9	9.48	0.3	CQ-vesicular sounds
15	9.32	0.35	wheezes	15	9.83	0.27	wheezes
Chart				**Chart**			
6	11	0.22	Rx with lab values	6	10.11	0.25	Rx with lab values
13	10.29	0.24	med/lab values	13	10.17	0.25	lab values/tx
19	11.1	0.25	med & lab values/VS	19	9.59	0.3	med & lab value

As can be seen from Table 1, Pairs of RN Alternate and Multiple-Choice Items by Type, the RN participants used higher ordering thinking skills (as evaluated by raters based on their talk alouds) more often for the ordered response, multiple response, clinical scenario/chart, and hot spot items as compared to the paired multiple-choice items. (Larger "measures" reflect increased levels of cognitive processing. The measure reflects a scaled score centered at 10). There were some items within the item types mentioned where the difference was not significant and could be considered the same level of cognitive processing. However, on average for every item type for RNs, except audio, there was a significant difference in cognitive processing as noted by increased measures for those items. For the audio items there was no difference from the paired multiple-choice items. There was a significant difference between the cognitive processing used to answer the alternate items as compared to the multiple-choice items. When the item pairs by item type are reviewed as in Figure 3, it can be shown that the chart/clinical scenario items require the most cognitive processing, the ordered response and multiple response the next highest level of cognitive processing, and audio and hot spot required the least cognitive processing based on the

talk alouds. It should be noted that the computer interface for Ordered Response item may add a level of complexity to the task of responding. That is, the necessity to track and "type in" the answer rather than 'drag and drop' to achieve sequencing may be making the item unnecessarily difficult either in terms of cognitive processing or item difficulty.

Figure 3. Cognitive Processing RN



It seems that if higher levels of cognitive processing can be equated with critical thinking, then the use of the alternate items—especially the clinical situation/chart format—may provide increased opportunity to assess this aspect of nursing competence.

For PN participants, a pattern similar to the RN results emerges. As can be seen by Table 2, Pairs of PN Alternate and Multiple-Choice Items by Type, when the level of cognitive processing as measured by raters evaluating talk aloud answers to paired alternate and multiple-choice items, there is a general increase in cognitive processing for the alternate items.

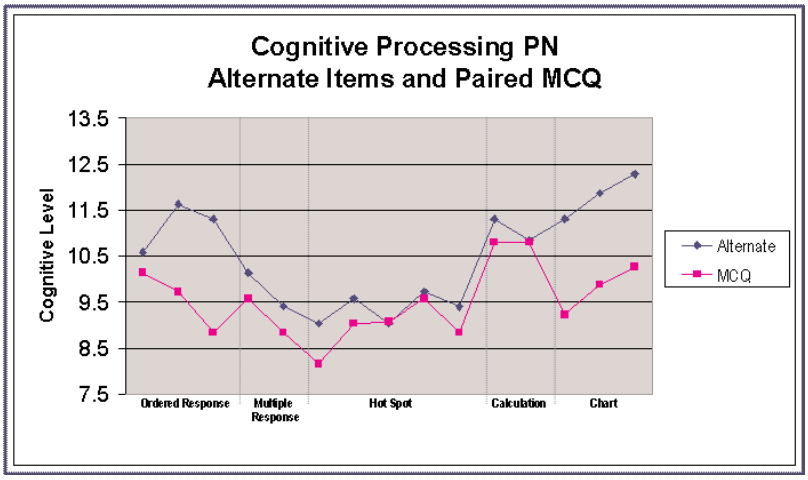
Table 2: Comparison of PN Alternate and Multiple Choice Items

PN							
Alternate Items				Multiple-Choice Items			
	Measure	SE	Label		Measure	SE	Label
Ordered Response				**Ordered Response**			
1	10.59	0.33	CPR adult	1	10.14	0.35	CPR adult
7	11.63	0.34	incision care	7	9.73	0.39	incision care
8	11.31	0.32	CPR-infant	8	8.84	0.46	CPR-infant
Multiple Response				**Multiple Response**			
10	10.14	0.35	G&D	10	9.58	0.4	G&D
13	9.41	0.42	infection control	13	8.84	0.46	infection control
14	9.04	0.45	reflexes	14	8.16	0.49	reflexes
Hot Spot				**Hot Spot**			
2	9.58	0.4	apical pulse	2	9.04	0.45	apical pulse
4	9.04	0.45	popliteal	4	9.07	0.46	popliteal
6	9.73	0.39	carotid pulse	6	9.58	0.4	pulses
16	9.4	0.43	dorsalis pedis	16	8.84	0.46	dorsalis pedis
Calculation				**Calculation**			
5	11.31	0.32	Calculation	5	10.8	0.32	calculation
12	10.85	0.33	Calculation-I&O	12	10.8	0.32	calculation
Chart				**Chart**			
3	11.31	0.32	lipitor/lab values	3	9.23	0.43	lipitor
9	11.87	0.35	med/lab values/VS/action	9	9.88	0.37	med
15	12.28	0.39	lab/meds/ASA	15	10.26	0.34	ASA
Fill-in-the-Blank				**Fill-in-the-Blank**			
11	10.01	0.36	Blank-pressure ulcer	11	9.73	0.39	pressure ulcer

As with the RN items, we note that within item types, some alternate items require the same amount of cognitive processing as multiple-choice items when the error associated with the measure for cognitive processing is considered. Thus for the hot spot and calculation items, some of the alternate items required the "same" level of processing as multiple-choice items. For the remainder of the items—ordered response, multiple response, and clinical scenario/chart—participants used more cognitive processing to answer the items.

In addition, for PNs in this study as seen in Figure 4, the clinical scenario/chart items are the most cognitively complex followed by the calculation, ordered response and multiple response items. The hot spot items were the least complex. It seems that if higher levels of cognitive processing can be equated with critical thinking, then the use of the alternate items—especially the clinical scenario/chart format—may be the wave of the future.

Figure 4. Cognitive Processing PN



Based on the results, raters are able to assess the cognitive processing that is used by participants to answer alternate and multiple-choice questions. In general, it appears that all of the alternate items except audio items require more complex cognitive processing when compared to a paired multiple-choice format items.

Conclusions

Experts are able to agree on ratings of examinees' cognitive processing based on the talk alouds of alternate and multiple-choice item types. Based on this study, experts should be able to categorize the cognitive processing that is used by candidates to answer examination items of various types. Historically, assumptions (based on the literature and expert judgment) have been made about how NCLEX® candidates process information in order to answer multiple-choice items. Now there is some empirical data to support the taxonomy used to categorize the cognitive processing of NCLEX items. It should be noted, however, that more research is needed to determine the cognitive processes and thus coding of the Create (synthesis) items since there were no items in this study that were targeted for the Create cognitive level. Additionally, results may not extend to other professions, and generalizability, even within nursing, may be an issue due to small sample size.

In this study, higher cognitive processing is required to answer many of the alternate item types when compared to the paired multiple-choice item. With the importance placed on these higher cognitive processes (i.e., critical thinking) by many educators, the use of alternate items may allow better assessment of the use of those higher cognitive levels. Perhaps the best strategy for assessing competence is to take advantage of a variety of item types to maximize the benefits of each item and thereby reduce the risk of using just one type of item on a high stakes examination. The results of this research provide test developers of high stakes examinations, whether they be educators or regulators, a comparative framework for an alternate path of item development aiming for a negotiated middle ground somewhere between austere and efficient multiple choice items and the stimulus rich, yet often unattainable simulation testing environment.

References

Anderson, L., Krathwold, D. (Eds.) (2001). *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives.* White Plains, NY: Addison Wesley Longman, Inc. 287-294.

Bloom, Benjamin. (1956). Major categories in the taxonomy of educational objectives. Retrieved March 25, 2004, from http://faculty.washington.edu/krumme/guides/bloom.html

Buckwalter, J., Schumacher, R., Albright, J. & Cooper, R. (1981). Use of an educational taxonomy for evaluation of cognitive performance. *Journal of Medical Education, Feb, 56*(2), 115-21.

Ericsson, K. A. and Simon, H. A. (1984). *Protocol Analysis: Verbal Reports as Data.* Boston, MA: MIT Press

Facione, N., Facione, P. & Sanchez, C. (1994). Critical thinking disposition as a measure of competent clinical judgment: the development of the California Critical Thinking Disposition Inventory. *Journal of Nursing Education, 33*, 345-350.

Haladyna, Thomas M. (1997). *Writing test items to evaluate higher order thinking.* Needham Heights, MA: Allyn & Bacon.

Katalin, Elekes. (2002). "Please keep talking": The 'think-aloud' method. Novelty, 7(3), Retrieved August 12, 2002, from http://www.novelty.hu/htm2/vol73/elekes.htm

Linacre, J.M. (2003). *A Users guide to FACETS No 3.41.0:* Rasch measurement computer program, Chicago, IL: MESA Press.

McDonald, M. (2002). *Systematic assessment of learning outcomes: Developing multiple-choice exams.* Sudbury, MA: Jones & Bartlett.

Miller, D., Sadler, J., Mohl, P. & Melchiode, G. (1991). The cognitive context of examinations in Psychiatry using Bloom's taxonomy. *Medical Education, Nov, 25*(6), 480-4.

Osterlind, S. & Merz, W. (1994). Building a taxonomy for constructed-response test items. *Educational Assessment,* 2(2), 133-148.

Parshall, C., Davey, T. & Pashley, P. (2000). Innovative item types for computerized testing in *Computerized Adaptive Testing Theory and Practice*. Van der Linden, W. & Glas, C. (Ed.) Netherlands: Kluwer Academic Publishers.

Taylor, C. (2000). Clinical problem solving in nursing: insights from the literature. *Journal of Advanced Nursing. Apr, 31*(4), 842-9.

White, A. (2003). Clinical decision making among fourth-year nursing students: an interpretive study. *Journal of Nursing Education.Vol. 42*, No.3. p 113-120.