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

From the Editors
1
F. Jay Breyer, Ph.D.
Michael Rosenfeld, Ph.D.

Columns
Abstracts and Updates
2
George T. Gray, Ed.D.

Technology and Testing
7
Robert C. Shaw, Jr., Ph.D.

Legal Beat
10
Dale J. Atkinson

Articles
Making the Test Development Process More Efficient Using Web-Based Virtual Meetings
13
Deborah L. Schnipke, Ph.D.
Kirk A. Becker

Assessing Critical Thinking Using a Talk-Aloud Protocol
18
Anne Wendt, Ph.D., RN, CAE
Lorraine E. Kenny, MS, RN
Casey Marks, Ph.D.
Welcome to the Winter 2007 issue of the CLEAR Exam Review. We have three columns and two articles that we think you will find interesting and informative.

George Gray, in “Abstracts and Updates,” cites articles dealing with the following topics: validity in court, point counter-point regarding task analysis methodology, point counter-point regarding standard-setting methodology, additional standard setting articles, item response theory, and differential item functioning.

Robert Shaw, in “Technology and Testing,” deals with the topic of web conferencing. He describes the primary function of a web conference, the features available, costs, security, and human factors issues. The column provides a good deal of useful information for those of you considering web conferencing.

Dale Atkinson’s legal column discusses a case involving a test preparation company and the procedures they used to make their practice examinations look like the actual licensing examination. These procedures included having members of the test preparation company take the licensing examination multiple times and write down questions on “scratch paper.” The court ruled that the test preparation company had violated copyright laws and ordered that they pay a substantial fine.

This issue also contains two feature articles. The first is written by Deborah L. Schnipke and Kirk A. Becker and deals with using web-based virtual meetings to make the test development process more efficient and cost-effective. It discusses the types and requirements for virtual meetings as well as their advantages and disadvantages.

The second article written by Anne Wendt, Lorraine Kenny, and Casey Marks describes the use of talk-aloud protocols to explore the kinds of cognitive functioning examinees – recently licensed people – use to answer both multiple-choice and alternate item types. They provide some interesting results and conclude that talk-aloud studies such as those described are feasible and provide further evidence of what certain items are measuring.

As always, if you have an idea for an article or topic you feel would interest CER readers, please send it to either of us at the addresses on the facing page.
Validity in Court


Sireci and Parker review a number of cases in which validity issues have been introduced in court and compare the court’s interpretation of validity to the interpretation of validity evidence provided in the Standards (1999). The authors note that “it is difficult to challenge a test for poor psychometric quality. Rather tests are most often challenged because of differential results between legally protected ‘minority’ groups of examinees and the ‘majority’ group….It is typically only after a test is challenged as being unfair with respect to disparate impact that its technical characteristics are subject to scrutiny.” (p. 29) Routes to challenging a test in court are Title VI or Title VII of the Civil Rights Act of 1964, and the Equal Protection or Due Process Clauses of the Fourteenth Amendment.

The authors present four cases, three that have been heard since the most recent Standards were published in 1999: Debra P. v. Turlington, GI Forum v. Texas Education Agency, Williams v. Ford Motor Company, and Hearn v. City of Jackson.

For the case of Debra P. v. Turlington, “In 1979, the state of Florida required students to pass a basic skills test to earn a high school diploma. Although failing students were allowed to retake the test, failure rates for African American students hovered around 20%, while those for Euro-American students were around 2%.” The case was heard in Federal District Court, appealed, and “the appeals court sent the case back to the District Court to evaluate the ‘curricular validity’ of the test. … The court concluded that the state had met its burden that the tested material was taught and sanctioned its use as a requirement for high school graduation.” (p. 30) Additionally, a challenge that students had not been given opportunity to prepare for the test was rejected by the court.

“GI Forum v. Texas Education Agency (1999) was similar to the Debra P. case in that it challenged Texas’ use of the Texas Assessment of Academic Skills exam as a high school graduation requirement. …Also like Debra P., the key validity arguments turned on ‘notice’ (time to prepare for the new test) and curricular validity.” (p. 30) In this case, “Four of the five sources of validity evidence named in the Standards were used: evidence based on test content, internal structure, relations to other variables, and consequences of testing...The GI Forum case provides an excellent example of how a comprehensive validity argument can lead to the successful defense of a test for use as a high school graduation requirement when challenged in court.” (p. 31)

Williams v. Ford Motor Company was an employment test for unskilled workers. Again the argument of the plaintiff was on the basis of disparate impact. The test was based on a job analysis, and a predictive validity study using on-the-job performance was conducted that involved approximately half minority workers in the study sample. The plaintiff was not successful in this case; however, Sireci and Parker observe that “the court may expect evidence of construct validity in cases where the connection between the content measured and the skills required for successful job performance is not well-established.” (p. 32)

“Hearn v. City of Jackson (2003) was another suit against an employment test based on disparate impact against African American police officers seeking promotion to sergeant.” (p. 32) The defense of the test was

George T. Gray, Ed.D., is Director of Test Development, Professional Development Solutions, ACT, Inc.
based on a job analysis and a content relevance study. The latter study consisted of review of test items by a panel of three sergeants who served as content experts. Despite the plaintiffs’ argument that a study demonstrating that a correlation with job performance was necessary, the court “concluded that the content validity that had been established was sufficient for demonstrating that the exam was appropriate for its intended use.” (p. 32)

The authors note that while the Standards take a comprehensive approach to the validity evidence, the courts take a “very practical view and look for evidence that test specifications were sensibly developed, test content is congruent with testing purpose, and issues of potential bias are considered throughout the test development process.” (p. 33)

Point Counter-Point #1 Task Analysis and KSAs


A previous issue of this column referenced the original article in which Wang, Schnipke, and Witt offered a framework for translating the knowledge, skill, and ability statements (KSAs) required to perform job-related tasks into a content outline (Educational Measurement: Issues and Practice 25(1), 15-22). Wang, Schnipke, and Witt also compared content outlines for a nurse aide examination derived directly from job tasks with a content outline based on KSAs and suggested that the version based on KSAs could be more precise and more justifiable. A recent issue of Educational Measurement: Issues and Practice carried comment on this article by LaDuca of the National Board of Medical Examiners and a response by Wang, Schnipke, and Witt.

LaDuca’s main criticism is that the authors “are insufficiently skeptical of the utility of the task inventory approach, especially in applications with complex professions.” (p. 33) He stated that “the task inventory method may not be appropriate for the professions because of the nature of the work, arising from the social construction of the professions (i.e., its theory), and the professional autonomy vested in the incumbents.” The professions such as medicine or law have autonomy. “Doctors are powerful people in their work settings because they have significant latitude to define the work, both for themselves and for many other persons…” (p. 32)

Wang, Schnipke, and Witt suggested that there is little disagreement between LaDuca’s perspective and their own. “Our paper is intended to address job analysis issues over a broader and diverse array of professions/occupations/fields that license and certify their practitioners, and our perspective is not limited to the context of credentialing physicians. For many such professions/occupations/fields, task inventory constitutes an appropriate approach to job analysis leading to the development of a content outline for the credentialing examination.” (p. 37) They also make the case that “ratings provided by human judges will never be perfectly reliable, but using fallible human judgments is better than collecting no opinions from those practicing in the profession/occupation/field.” (p. 36)

Point Counter-Point #2 Standard-Setting Methodology


Reckase provides a framework for thinking about standard setting “that borrows some of the conceptual framework from true score theory and simulation procedures from item response theory.” (p. 4) “The theory is that panelists in a standard setting activity have an unobservable ICS (intended cut score), and it is the goal of the standard setting process to obtain an estimate of the ICS. Similar to true score theory, it is hypothesized that the standard setting process can be replicated and the distribution of replicated estimates of cut scores can be used to evaluate the quality of the standard setting process.” (p. 12) Simulations were conducted for both the modified Angoff and bookmark procedures, using both the Rasch model and the three-parameter logistic model.

As part of the discussion section of the paper, Reckase observed that “overall, the proposed theory provides a framework for thinking about standard setting processes and for evaluating the procedures using the proposed criteria…. (t)he modified Angoff method functioned better when proportion-correct ratings to two decimal places were used. Rounding the ratings to one decimal place resulted in poorer recovery of the intended standard, especially at the extremes of the proficiency range. …The evaluation of the bookmark method indicated that averaging the locations of the items on either side of the bookmark gave better results

Schultz praised Reckase for developing the concept of the panelist’s intended cut score; however, he “hoped to show that Reckase’s conceptual framework is too simplistic, that his assumptions for the Angoff method is unrealistic, and that his assumptions about the bookmark method do not capture what the panelists are actually doing. Because of these problems, Dr. Reckase’s simulations have little value for interpreting the actual outcomes of bookmark and modified Angoff procedures.” (p. 4) Schultz argues that different methods might be expected to produce different cut scores. The ordered item book used in the bookmark method, for example, provides an organizational structure for that method that is not present in the Angoff method. Another possibility is that the use of different response probabilities for the bookmark method may be associated with differences between Angoff and bookmark results. Schulz also suggests that the intended cut scores of individuals may be of less interest than the standard error of a mean cut score to “take into account the dependence that begins developing among panelists within a meeting when the Round 1 mean or median is presented as feedback.” (p. 5) Finally, the Reckase comparison of standard setting to psychometrics breaks down because standard setting involves two data sets: data from the test and data from the judges.

In contrast to Reckase’s simulation, Schultz describes a “real data” project to set achievement levels for the 2005 NAEP (National Assessment of Educational Progress) Grade 12 mathematics test using the Mapmark method, a bookmark method that uses item mapping (Schultz and Mitzel, 2005). He concludes that “bookmark placement cannot be predicted from real item rating data.” (p. 7) In addition, he considers the effects of rater training and multiple rounds of rating, not just one round of initial rating.


Reckase’s response compares his concept of standard setting to that of Schultz and shows how the models of error suggested by Schultz can be used to evaluate standard setting methods using the procedures described in Reckase’s original article. Reckase states that “the different error models that are proposed make it very clear that there is not enough research on the mental processes used by standard setting panelists and the effects of the errors in judgment on the estimates of standards that result from different standard setting methods.” (p. 17) He also concludes that “the modification to the bookmark method recommended by Schultz, and the encouragement of panelists to consider a range of items, shows considerable promise over the method based on selecting the first item judged to have a probability of correct response below the mapping probability.” (p. 17)

Additional Standard Setting Articles


The Bookmark and Item Mapping methods are both standard setting methods in which item response theory is used to enable a systematic translation of professional judgments into a passing score. One aspect of this process (in either case) that is often overlooked is a measurement decision that figures into the result of the process. For both methods, the goal is typically to identify the point at which the minimally proficient candidate has a 67% probability of answering an item correctly. Of necessity, some value must be selected. Sixty seven percent is most frequently used as this value, but other choices have sometimes been made. A different response probability may result in a different passing score. In this paper, Huynh explains the rationale for the use of 67% as the response probability criterion.

Huynh notes that response probability values ranging from .50 to .80 have been used for various purposes. Huynh first proposed the use of .67 for a Rasch binary item in 1994 and subsequently extended this research to other IRT models. Huynh indicates that the focus of his work has been on information of the correct response, not the total item information or test information. “Because the probability of the correct response is p, the proportion of the total item information partitioned to the correct response is taken as p(1-p)p. This information is maximized when p = 2/3 or .67.” (p. 20)

This is an article that merits careful reading. Many papers that are concerned with standard setting refer to the selec-
tion of a response probability, including the Reckase and Schultz papers that are presented above.


The study reported here investigated the impact of judge characteristics on the standards established by two health care certification boards. In both cases a panel of approximately 30 judges used the Rasch-based Objective Standard Setting method which was developed by Stone. The procedure did not include an opportunity for judges to discuss and revise their ratings, as Stone wished to preserve the characteristics of the judges’ own original ratings. Seven judge panel characteristics were included in the analysis: gender, ethnicity (4 classifications), geographical regions (4), number of years certified (3 levels), practitioner or academic environment, practice setting (private, clinic or hospital based), and private pay or public pay. Analysis using Rasch FACETS showed no “significant differences” for gender or ethnicity for either panel group, but the other five variables “were all observed to influence the standards established.” (p. 164) Stone concludes that “…the usefulness and success of the methodology selected may depend as much upon the selection of expert judges used in the process as on the selection of the model itself.” (p. 165)

**Item Response Theory**


Linacre states that “rank-ordered data have characteristics that align well with the family of Rasch measurement models…. If there are two or more rankings of the same elements, then there may be enough information to construct interval measures of the distances between the elements.” (p. 130) He explores this possibility using data from 47 USPGA golf tournaments in the 2004 season. Four analytic approaches were described, two independent paired comparison approaches, the dependent paired-comparison approach, and the rank order approach. Each of these has advantages and disadvantages and can be implemented using the WINSTEPS and FACETS Rasch analysis programs.


This paper compares equating using the characteristic curves for the correct answers of items with a method that considers the characteristic curves for all of the item’s options. Studies using simulated and real data were conducted. The author concludes that equating using all distractors is “quite promising in support of the use of distractor categories for reducing the number of linking items needed to achieve good linking precision.” This advantage appears to be greatest where the number of linking items is small and decreases as the number of linking items is increased to ten or greater. (p. 209)

**Differential Item Functioning**


This article explores differential test functioning (DFT), “the aggregated DIF effect across the items of a test or a subset of items.” (p. 295) The authors note that DFT is important because of the aggregate effect of items with relatively little DIF, or, conversely, the possibility that an item with essential content showing DIF may have little impact on DFT. The authors investigate the use of two generalized DIF effect variance estimators, an unweighted estimator and a weighted estimator. They conclude that “because these estimators measure the total spread among the item-level DIF effects, both positive and negative, they serve as unsigned measures of DFT.” (p. 309)


This study explored the use of three statistical tests for use with a graphical analysis of DIF in a simulation study and a study comparing translated versions of a high school certification examination. Graphic analysis comparing item response functions was performed using TESTGRAF Software. Statistical analysis was conducted using Fisher’s $X^2$ test, Cochran’s Z test, and Goodman’s U test. A regression correction procedure was used to reduce bias in
comparing item response functions. The authors discuss the situations in which each of these tests might be applied most appropriately.


The authors compared four methods for detecting differential item functioning (DIF) in ordinal items: the Mantel, generalized Mantel-Haenszel (GMH), logistic discriminant function analysis (LDFA), and unconstrained cumulative logits ordinal regression (UCLOLR). The authors concluded that “all procedures had good Type I error control, as well as high power for detecting uniform DIF. However, the Mantel could not detect nonuniform DIF, and the LDFA also performed poorly in detecting nonuniform DIF, particularly when item discrimination was high. The UCLOLR and GMH performed extremely well under conditions simulated in this study.” (p. 935)

**References Cited**


If psychometricians or regulatory agencies could write their own certification tests, then test development would be easier than it is. However, this practice would usually punch a wide hole in evidence supporting valid testing outcomes. We rely on collaboration among a group of experts to build validity evidence at several points in the process. These experts typically travel to a location where they meet for a day or more, eat too much for lunch, and make critical decisions about the test. These meetings consume significant resources, so web conferencing alternatives have grown as a way to facilitate work while saving travel expenses. I have used these technologies for the last four years to develop certification tests, so I thought I would touch on some points that might help someone who was considering an alternative to traditional meetings.

Choosing Among Vendors and Features
Web conferencing is a fairly new service on the business landscape. There has not been a period yet when the industry consolidates into a few choices. I found a useful resource at http://www.conferzone.com/ that helps sort through dozens of web conferencing vendors from which one can choose. The site includes reviews of several vendors which an organization might find useful.

The primary function of a web conference is to share content with a group. Some content is prepared before a meeting and presented in a training format, like an item writing workshop or a passing point study. Other content will change while it is shared like test items or an inventory of tasks for a job analysis.

Most vendors have evolved elaborate systems that supplement basic functionality. An organization should consider the utility of these features before choosing a system. A quick list of additional features follows:

- Brand a meeting site with an organization’s name and logo
- Dedicate a toll-free number to all of an organization’s meetings
- Integrate telephone conferencing into the Internet interface under the control of a meeting host
- Permit interactions among participants and hosts with tools like polling, chat, hand raising, and notes
- Provide an exclusive meeting space and ID for each host (organizational employee)
- Record a session
- Schedule meetings without a reservation
- Share a whiteboard for brainstorming
- Transfer encrypted files securely

Robert C. Shaw, Jr., Ph.D. is a Program Director for Applied Measurement Professionals, Inc.
Cost
Web conferencing is a volume-based business. The more meetings an organization hosts, the more it will spend. An organization should estimate the number of people who will host meetings along with average and maximum numbers of participants for its meetings.

Some vendors require contractual commitments, while others do not require contracts. There are two pricing models offered by most vendors: (1) pay-per-use and (2) pay-per-seat. The more you buy, the lower the rate per minute, which particularly applies to the pay-per-seat model.

Organizations that select the pay-per-seat model may eventually face overage charges. As employees of the organization and meeting participants grow more comfortable with web conferencing, overlapping meetings become more common. Sometimes an organization’s seat limit is exceeded, for which some vendors will happily submit a premium charge.

Some vendors may require a one-time set-up charge, particularly if an organization’s meeting site will be branded. An important supplement to branding should be selection of a simple, recognizable URL address so users can easily find their way to meetings using an Internet browser.

Lastly, I encourage negotiation with potential vendors on services and prices. A vendor may be willing to waive the set-up fee to earn your business for a couple of years. The rate per seat or user can be negotiated along with the telephone conferencing rate.

Security
Because files shared during a web conference remain on the computer of a host, meeting content is as temporary as a telephone call on the vendor’s server. If a document is changed on the host’s computer, the next image sent to conference participants will show the change. As long as the Internet connection stays on and page images refresh every few milliseconds, document changes can appear quite fluid. However, documents never exist on participants’ computers or even the server of the web conferencing vendor.

Like a telephone call, it would be technologically possible for a motivated hacker to tap into a session. However, this risk is no greater than the risk of taking a computer with an item bank out into the world to facilitate a face-to-face meeting.

Hacking into a test development session would require a high degree of technological skill to accomplish without detection by the vendor. Less sophisticated individuals could try to join a web conference meeting to which they did not belong. However, vendors’ systems typically authenticate each participant with a meeting password.

Some meeting participants may use computers owned by their employer. An increasing number of organizations have increased internal security to prevent Internet-based instructions and ports from operating. These prevention measures can stop installation of a web conferencing vendor’s application or browser plug-in, which effectively stops a participant from joining a meeting. I typically advise users to involve their local technology staff to intervene and permit the installation of web-conferencing components. This obstacle can be overcome with enough lead time.

Human Factors
Personal experience tells me human factors are the most difficult obstacles to successful web conferences. People resist change and can be intimidated by the technology. I will host passive practice sessions in the days leading up to a web conference of new users. Doing so gives participants an opportunity to apply their new-found skills in a low-stakes environment without time pressure.

Anticipate some experts will argue that their interactions and the test will suffer when developed by web conferencing, particularly when they have previously participated in face-to-face meetings. My observations indicate interactions of experts can be just as vigorous, particularly when participants already know one another. It is true that a meeting facilitator is unable to detect negative body language among those who will not vocalize disagreement until it is drawn out of them. However, a solution is to adjust procedures to elicit a response from every participant rather than assume that silence signals agreement.

Scheduling web conference meetings can be difficult. We often start by asking participants when they are available. However, web conferences tend to become events inserted into the busy calendars of expert participants. While participants are accustomed to taking days off from work to attend a face-to-face meeting, I have observed that they typically will not do the same for a web conference. An effective technique is to set the date and time for a meeting several weeks in advance and populate the meeting with experts who can accommodate the meeting in their schedules.

One of the truly powerful dimensions of web conferencing is the capacity to bring experts together from around the globe who would otherwise have great difficulty meeting. However, finding a meeting time can be a challenge when these participants reside in widely different time zones. A resource that helps me sort out time zones can be accessed at http://www.timezoneconverter.com/cgi-bin/tzc.tzc.
Summary
Web conferencing is a technology that permits sufficient interactions among experts to conduct test development activities. If substituted for face-to-face meetings, anticipate it will be more complicated to schedule a meeting. The payoff is that resources expended to host meetings are usually reduced. There are several vendors and services from which an organization can select. The security of test materials during these meetings should be as high as it is for face-to-face meetings.

References

The development, administration and maintenance of a defensible, uniform, high stakes licensure examination program used by governmental entities or private sector certification programs involves many complex and costly processes. Examination owners undertake numerous measures to ensure the integrity and security of such examinations for the benefit of those that rely upon the results in making essential public protection decisions related to licensure (or denial) of applicants. Under certain circumstances, these entry level, minimum competence determiners may also be used as part of a reinstatement of a disciplined licensee or one who has been removed from the practice for an extended period of time.

As technology evolves and exam administration modalities advance, so do ways to circumvent the exam administration processes, creating the need to continually proctor examinees. It appears that attitudes condoning “cheating” on educational examinations have also evolved almost to the point of acceptance in society. Examination owners must take all necessary steps to protect the security of their programs from candidates using advanced technology to cheat. However, sometimes good old fashioned scratch paper used in an exam program can facilitate exam breaches. As an illustration of the “damage” to an exam program that can result from a security breach, consider the following.

The National Conference of Bar Examiners (NCBE) develops testing materials used by more than 50 jurisdictions to evaluate applicants seeking admission to the bar. The Multistate Bar Examination (MBE) is one of the most widely used NCBE products. The MBE is a 200-question multiple choice examination administered twice per year. One criterion for admission to the bar in most jurisdictions is to successfully complete the MBE and a separate essay.

As with any of the uniform licensure examinations, the development, administration and maintenance of a minimum competence assessment mechanism is an involved process. The MBE reuses many questions from one administration to the next, thus the NCBE undertakes many steps to maintain the secrecy of its item banks, including submission of the items to the United States Registrar of Copyrights and prohibiting test takers from discussing or reproducing MBE questions. The NCBE has also pursued legal action against bar review courses that violate the copyright protections.

The Multistate Legal Studies, Inc. (known as the Preliminary Multistate Bar Review “PMBR”) is a test preparation company whose purpose is to prepare students for the MBE. The PMBR offers a variety of preparation programs including a three day class, a six day class, and one-on-one tutorials. The courses provide oral and written instructional materials addressing the substantive law tested on the MBE, as well as test-taking strategies and practice MBE questions. Advertisements for the company use “testimonials” from former students emphasizing the similarity of the PMBR practice questions to those actually found on the MBE. The PMBR is taken by approximately 60% of MBE examinees and produced over $16 million dollars in revenue in 2004 and over $35 million dollars in the years 2001 through 2004. The two owners of the company have drawn millions of dollars in salaries over that time frame.

PMBR students take a complete simulated MBE (referred to as the PMBE) and, thereafter, receive instruction on substantive areas

Dale J. Atkinson, Esq. is a Partner in the law firm of Atkinson & Atkinson, which represents a number of boards. He received his law degree from Northwestern School of Law, Portland, OR.

“As technology evolves and exam administration modalities advance, so do ways to circumvent the exam administration processes, creating the need to continually proctor examinees.”

Dale J. Atkinson, Esq. is a Partner in the law firm of Atkinson & Atkinson, which represents a number of boards. He received his law degree from Northwestern School of Law, Portland, OR.
of the law and test taking techniques. Students also receive written answer keys to the PBME with detailed explanations and citations to source materials used to develop the questions. Owner one (Mr. F) generated almost all of the PMBE questions and answer keys himself relying on hornbooks, treatises, reported and published cases. However, Mr. F also took the MBE more than 20 times, including taking the bar examinations in Alaska (which includes the MBE) on five occasions, barely passing it on the 5th attempt. Mr. F admits that he used notes of other PMBR employees who have sat for the MBE on numerous occasions.

Because her scores on the MBE were so poor (the exams are cumulated to determine passing scores), owner two (Ms. Z) twice failed the Kentucky Bar Exam, in spite of passing the essay portion. All in all, Ms. Z took the MBE (Ms. Z) twice failed the Kentucky Bar Exam, in spite of passing the essay portion. All in all, Ms. Z took the MBE (which includes the MBE) on five occasions, barely passing it on the 5th attempt. Mr. F admits that he used notes of other PMBR employees who have sat for the MBE on numerous occasions.

In February 2003, Mr. F took the MBE in Anchorage, Alaska, the only jurisdiction that allowed the use of scratch paper, and was caught attempting to remove the scratch paper from the test center in violation of examination policies. The exam proctor confiscated the notes and filed an irregularity report. Based upon the irregularity report, the NCBE launched an investigation extensively reviewing the PMBR course materials and comparing PMBE questions with those actually contained on the MBE. After concluding that more than 100 items had been copied, NCBE filed litigation in the United States District Court Eastern District Pennsylvania against the Multistate Legal Studies, Inc. (referred to as defendant) alleging copyright infringement. The matter proceeded to a bench trial (non-jury) which lasted several days.

In a copyright infringement action, plaintiffs must prove both ownership and copying of original elements of the protected works. Based upon registrations with the Copyright Office and other factors, the court held that the NCBE “owned” the protected work. Thus, the only remaining question was whether the protected works were copied.

In assessing the evidentiary matters, the court outlined the actions of the defendant. It noted that Mr. F and other PMBR employees regularly wrote down information about the fact patterns, prompts and answer choices appearing on the MBE that they had taken. In order to facilitate this process, Mr. F and other PMBR employees gravitated to the only jurisdiction that allowed the use of scratch paper.

The court also noted that the promotional brochures and testimonials “brag” about how close its questions are to those on the MBE. In addition, Mr. F, in actually conducting the courses, also emphasized the similarities between such questions. Finally, the evidence established that many of the PMBE questions were near, if not exactly, verbatim to the MBE items, including distractors and answer keys. Based upon the evidence, the court concluded that the defendant had willfully copied MBE questions “...either by setting out to do so, or engaging in behavior that was so certain to lead to copying that intent must be inferred.”

The court engaged in an analysis that compared many items from the MBE to the review materials to illustrate its findings. The similarities and, at times, verbatim use “...practically leaped from the page.” In addition to these egregious examples of infringement, the court also opined that “...less than wholesale reproduction can also provide a sufficient basis to conclude that there was copying.” The court pointed out examples of these less than wholesale reproductions.

Having determined that there was copying, the court addressed the issue of whether the copied elements are subject to copyright protection. The defendant argued that the works should be afforded only limited protection because they “…tested established legal rules within a relatively narrow set of formal constraints.” In rejecting this argument, the court held that teaching legal principles tested on the MBE is permissible, but doing so through the use of the same fact patterns, prompts, and answer choice combinations contained on the MBE is not permissible. Noting prohibitions under the copyright laws, the court cited examples of PMBE items containing similar fact patterns with four substantially identical answer choices to the MBE.

Further, the defendant argued that its work was independently created, citing “source binders” identifying the research for the PMBE items. The court found this argument to be “wholly incredible” stating that the cited sources simply provided support for the legal principle being tested, but lacked in any evidence as to independent creation. Indeed, the court noted that the source binders submitted by defendant were “…simply ad hoc efforts to identify sources that could, theoretically, have been used.” Further, the court emphasized that in the 18 months after agreeing not to take the MBE pending the outcome of the litigation, the defendant did not draft any new PMBE questions.
Finally, the court rejected the arguments by the defendant of laches and estoppel. [Editors’ note: Laches and estoppel are two different legal principles that essentially amount to a waiver of legal rights by failing to assert them in a timely manner. By not asserting the claim and engaging in contrary conduct, the party is stopped from asserting the associated legal rights. A practical example would be claiming a credit card charge is fraudulent after paying the charge.] The court held that the NCBE diligently pursued the litigation upon notice of potential violations and that the defendant could not have reasonably taken any silence after publication as general permission to engage in copyright infringement.

Based upon its findings, the court examined the damages award to which the NCBE was entitled. The court reviewed the financial aspects of the defendant’s business which amounted to over $35 million dollars for the period of 2001 through 2004. It held that the egregious nature of the defendant’s actions and the marketing strategies of promoting the similarities between the PMBE and the MBE warranted an award to the NCBE of $11,902,787, one-third of the defendant’s revenues. The court also enjoined the owners and other PMBR employees from taking the MBE, except for purposes of seeking licensure in the state in which they took the exam. Finally, the court awarded costs and attorneys fees to the NCBE.

The District Court, applying the copyright laws, recognized the importance of the licensure examinations, noting the:

“Defendants’ willful and egregious copyright infringement harmed the public as well as the plaintiff. States have a compelling interest in regulating admission to the bar both to maintain the integrity of the legal system and to protect the safety of their citizens. By exposing its students to questions likely to appear on the MBE, PMBR undermined the integrity of the bar examination, possibly causing the admission of unqualified applicants.”

National Conference of Bar Examiners v. Multistate Legal Studies, Inc. 2006 WL 2460903 (District Court PA 2006)
Making the Test Development Process More Efficient Using Web-Based Virtual Meetings

Deborah L. Schnipke, Ph.D.
Kirk A. Becker

The test development process includes many steps, specifically the initial job task analysis, content outline development, item writing, test construction, field testing, item analysis, standard setting, and many others. Many steps require the input of a variety of Subject Matter Experts (SMEs) who may be located throughout the country or internationally. In the past, a variety of in-person test development meetings were required. We have found that we can make the test development process more efficient by having web-based virtual meetings, either in addition to, or instead of, in-person meetings. In this paper, we review reasons for having virtual meetings, requirements for virtual meetings, advantages and disadvantages of virtual meetings, and guidelines for success. Our goal in this paper is not to promote any particular web conferencing system, but rather to discuss general features that we find useful and to provide examples of the types of meetings we now sometimes run virtually.

Types of and Requirements for Virtual Meetings

The simplest type of virtual meeting is a teleconference. Some topics might require inclusion of documents which are provided to participants in advance for use during the meeting. The only requirement is for each participant to have a telephone line, and possibly a way to receive documents in advance (e.g., email, postal mail, or fax). These meetings can be very effective to discuss certain types of tasks and will likely continue. Teleconferences, however, do not provide a very rich experience for participants, and keeping them engaged and involved can be a challenge. Additionally, providing documents in advance may be a security risk.

Another medium for a virtual meeting is video teleconferencing. Video teleconferencing requires participants to have access to video teleconferencing equipment which is far more expensive and not nearly as readily available as a regular phone line. However, video teleconferencing provides a richer experience for participants because they can actually see each other, which provides body language information and other visual feedback. Video teleconferencing keeps the participants more engaged and interested and is recommended over non-video teleconferencing when possible. If participants are in more than two locations, however, video teleconferencing becomes less feasible.

Teleconferencing (with or without video) by itself is not sufficient for many of the types of test development meetings that we run. For example, some of the additional features we find useful are:
- share documents, applications, or entire desktop
- control what the participants are viewing at any point in time, while allowing participants to view materials asynchronously when appropriate
- enable participants to ask questions or otherwise communicate with the facilitator(s) without disrupting the meeting
- collect responses from SMEs anonymously

Web-based conferencing systems provide these features and more, and we have found web conferencing, in conjunction with teleconferencing, to be very effective for test development meetings.

For each type of meeting (e.g., training, item review, standard setting, job analysis, item writing), the appropriateness of doing a web-based virtual meeting and the features required need to be determined. Features to consider include:
- Bandwidth – does the web conferencing system provide these features and more, and we have found web conferencing, in conjunction with teleconferencing, to be very effective for test development meetings.

Deborah Schnipke is a Principal Consultant at Virtual Psychometrics and Kirk A. Becker is a Manager in Measurement Services at Pearson VUE.
system require a high-speed internet connection? Do all the participants have high-speed connections?

- Sharing applications – what types of material need to be shared? Word processor documents, spreadsheets, item banking applications, etc?
- Messaging (public/private) – does the web conferencing system provide a method for online messaging? Should participants be allowed to message each other, or just with the facilitator?
- Plug-ins – are plug-ins required to use the web conferencing system?
- Security – does the web conferencing system provide secure, password-protected access?
- Asynchronous viewing – when viewing documents, do participants need to be able to scroll on their own? Can this feature be enabled/disabled as needed?
- Firewall issues – will all participants be able to access the web conferencing system?

Types of Virtual Test Development Meetings

The authors have successfully used virtual meetings to run a variety of psychometric/test development meetings. This includes item review and development, standard setting, job analysis or content outline development, and training presentations. Since the purpose of this paper is not to advocate for a particular web conferencing system, we will describe general capabilities of this type of service and how to make use of these to run different types of test development meetings.

One of the main benefits of web-based conferencing is the ability to share content without distributing hard or soft copies of sensitive material. In our experience, the most critical additional features of a web conferencing system are capabilities for synchronous and asynchronous viewing of presented content, and the ability to allow chatting between SMEs and facilitators while restricting chatting between participants. Additional features might include whiteboarding, recording, integration of phone and video conferencing, closed captioning, and many more. We will describe these features relative to how they have been used to run different types of test development meetings.

Figure 1 shows a sample layout screen similar to several web conferencing systems that we use. In this example, the presenter displayed an item by sharing a Word document. The Word document is displayed in the large window on the left. The chat feature is available in the lower right window, and the list of participants is available in the upper right window.

Item review and development

One of the key concerns for item review and development is the security of the test content. While any item review session is only as secure as the SMEs participating, mailing or emailing test content provides significantly more risk than web-based conferences. Conducting an item review through a virtual meeting allows for the review of extremely large item banks in a short period of time. SMEs have reported that they enjoy the convenience of reviewing items from their own space, as well as the perception that they can better focus on the items. We have typically used virtual item review meetings with relatively small groups of SMEs to screen item banks and flag items for potential problems, and then review flagged items by larger groups of SMEs at an in-person meeting. Separating item review from item revision, with the exception of small changes to content, keeps SMEs focused on a single task. For virtual item reviews, we have found that it is useful for the facilitator(s) to keep a checklist with item ids, a check box for each SME, and a space for comments. The checklist allows the meeting to move along while simultaneously keeping track of all items reviewed and any possible problems.
Virtual meetings allow psychometricians to conveniently conduct standard setting meetings. In addition to the ability to present documents, standard setting may require that SMEs can asynchronously view presented documents, and that chat between SMEs and the host can be private. Asynchronous viewing allows SMEs to review and rate items at their own pace, which is useful for bookmark standard setting. When using the Modified Angoff procedure, we have found that facilitator-control of the display helps keep the ratings organized. Restricting chat allows SMEs to provide their item ratings to the host privately, so that initial ratings are independent.

As with in-person standard setting meetings, a virtual standard setting begins with a discussion of the characteristics of a minimally competent candidate. Relevant documents (e.g., the test content outline) as well as notes from this discussion are available through document sharing. SMEs are able to provide item ratings to the host privately through the chat function. Once all SMEs have rated the items, the follow-up discussion and revisions to ratings can take place. In some instances it may be beneficial to have SMEs drop off the phone conference while rating items and dial back in when they are finished. This option allows SMEs to ‘talk shop’ if some of them finish early.

Job analysis and content outline development

Like an in-person focus group, virtual meetings bring together SMEs in a single location to discuss a topic. A virtual meeting can therefore facilitate a job analysis by providing SMEs with a setting to define content areas, tasks, and areas of knowledge for use in the development of a test content outline. Virtual meetings can, depending on the composition of the participants, provide an opportunity for participants to either define the job analysis survey or generate weights for the content outline. The capability for either white boarding or sharing a word processing program allows participants in this type of meeting to see the job analysis survey or content outline as it is built. Some SMEs find the presentation on their computer screen more amenable to review than the traditional poster paper or projection screen used for in-person meetings.

Training presentations

Training presentations can be effectively given through virtual meetings or hybrid virtual/in-person settings. If participants in a virtual training are joining from countries with expensive or irregular phone service, the use of VOIP (voice over internet protocol) may be less expensive and more reliable. Some teleconferencing services allow for transmission only, which allows a presenter to speak uninterrupted to large groups of participants. Whether presenting only to virtual participants, or to both in-person and virtual attendees, the use of a co-presenter is also recommended. A co-presenter can respond to questions asked through the chat functionality, or relay questions to the presenter so that all participants can hear the question and the response.

Advantages of Virtual Meetings

In the test development process, we want to draw on the best people possible to serve as Subject Matter Experts. Frequently, these people are geographically distant. In-person meetings can be a challenge to coordinate. Not only is there the high cost of travel, but a bigger constraint is frequently the time involved in traveling to a common location. Not only do SMEs need to take time out of their busy schedules for the meeting itself, but they also need to factor in travel time, which is not viable for some SMEs. Coordinating everyone's schedules can be an issue, and if SMEs are traveling across time zones, they may also be affected by jet lag. Thus, holding a meeting but avoiding the travel is desired if feasible.

The cost of virtual meetings is typically much less than in-person meetings. Flight, lodging, and meeting room costs are all replaced by the lower cost of running the internet conferencing. To justify the expense of in-person meetings, multiple activities are frequently performed over the course of the meeting, while virtual meetings typically focus on a single activity. Because of the increased focus on a single activity, SME participants can be selected from a more narrowly focused area of expertise. Additionally, smaller testing programs may find that test development activities that would otherwise be outside of their budget because of on-site meeting costs can instead be performed using virtual meetings.

Although the lack of travel may be a disadvantage for SMEs who see travel as a perk, the increased convenience of virtual meetings is a definite advantage. SMEs can participate in virtual meetings from the office or home and only have to block out a limited time period. Even within a large centralized campus, this technology may provide greater convenience among participants. Because of the increased convenience and shorter time commitment, virtual meetings have faster scheduling and turnaround than in-person meetings. We have found that virtual meetings can
typically be scheduled within approximately 2 weeks, while in-person meetings typically require at least 1-2 months of lead time.

With shorter, more focused meetings, virtual meetings may require several groups of SMEs or separate meetings to finish a large task. While recruiting and organizing multiple small meetings may be more time-consuming than organizing a single long in-person meeting, virtual meetings can allow for a broader range of participants in terms of expertise and geographic representation. It is far less expensive to include participants from Alaska and Hawaii, for example, in a virtual meeting, and typically much easier to find willing participants. In other words, it is easier to recruit SMEs for a 4-hour virtual meeting than for a 3-day in-person meeting in another time zone.

The use of virtual meetings for training allows larger groups of people to access the material. Session recording, automated transcripts, and file sharing can easily disseminate the training materials and the session itself to participants. Virtual meetings may allow for near real-time sessions on scoring rubric issues for raters or the review of draft test items written by new item writers.

Disadvantages of Web-Based Virtual Meetings

Web-based virtual meetings offer a great deal of flexibility and cost saving, but they are not a panacea and there are potential problems with this technology. The first, and most basic issue, is getting a web-based virtual meeting up-and-running. Although vendors of web-based conferencing software have built their systems for easy access, issues such as participant’s firewall setup, operating system, administrative access, and hardware may make connecting to a web-based virtual meeting difficult or impossible. We have found that most SMEs have been able to participate, but not 100%. For smaller meetings in which every SME counts, we recommend at least a 15 minute lead-time to resolve any potential issues with connectivity, as well as having first-time participants access the system a day or more in advance to ensure that they can connect.

The facilitator needs to decide, in advance, on the format for each part of the presentation (self-directed vs. host-controlled). The introduction may be host-controlled, for example, followed by a self-directed review of items. All of the various software options need to be reviewed prior to the beginning of the meeting to make sure the settings reflect the requirements of the meeting (e.g., messaging enabled with host, "save" function disabled for participants).

The various web-based conferencing systems have many options and can be intimidating at first. Not everyone needs an in-depth knowledge of the software; however someone on the team should be familiar with software and available to help set up the meeting. It is also a good idea to do a dry run internally before running a meeting with external SMEs.

Tips for Running a Successful Remote Meeting

If possible, we ask new participants to log into a practice meeting before the day of the actual meeting to make sure they don't have any connection issues. We have had cases where someone's firewall prevented them from accessing the web conferencing software, and they needed to have tech support help them. Working out these issues before the actual meeting is important. This may not be practical with extremely large meetings (e.g., training sessions with dozens or hundreds of attendees).

The facilitator needs to decide, in advance, on the format for each part of the presentation (self-directed vs. host-controlled). The introduction may be host-controlled, for example, followed by a self-directed review of items. All of the various software options need to be reviewed prior to the beginning of the meeting to make sure the settings reflect the requirements of the meeting (e.g., messaging enabled with host, "save" function disabled for participants).

The various web-based conferencing systems have many options and can be intimidating at first. Not everyone needs an in-depth knowledge of the software; however someone on the team should be familiar with software and available to help set up the meeting. It is also a good idea to do a dry run internally before running a meeting with external SMEs.

Having multiple facilitators/hosts during the meeting is beneficial. As the main facilitator runs the meeting, it is useful to have a second host deal with participants individually in the messaging window to answer questions, send and receive comments, and so forth. This allows participants who need extra attention to receive it without disrupting the meeting. Depending on the characteristics of
the meeting and the participants, it may be beneficial for the second host to be able to troubleshoot access to the system, or to be knowledgeable about the content and able to answer content-related questions.

If any printed materials will be used by the facilitator and/or participants during the meeting, those materials need to be prepared and printed in advance. If the participants will be using printed materials, the facilitator will need to verify at the start of the meeting that they have the materials available.

**Conclusion**

We have found that virtual meetings are most effective as a supplement – not a replacement – for in-person meetings. When possible, we use virtual meetings to perform some tasks before an in-person meeting, allowing the in-person meeting to be more focused and productive. We also use virtual meetings to allow more frequent contact with SMEs, which can speed up the test development process, rather than waiting for the next in-person meeting.
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 & Krathwol, 2001). Indeed, some well-known authors have...
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

<table>
<thead>
<tr>
<th>Traditional Item Type</th>
<th>Alternate Item Type</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Multiple Choice Item:</strong> 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?</td>
<td></td>
</tr>
<tr>
<td>1. Place the client in a private room.</td>
<td></td>
</tr>
<tr>
<td>2. Ask the client to wear a surgical mask.</td>
<td></td>
</tr>
<tr>
<td>3. Use sterile gloves to remove the wound dressing.</td>
<td></td>
</tr>
<tr>
<td>4. Wear a protective gown when providing wound care.</td>
<td></td>
</tr>
<tr>
<td><strong>Fill in the Blank/Ordered Response:</strong> 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.):</td>
<td></td>
</tr>
<tr>
<td>_____ test for rebound tenderness</td>
<td></td>
</tr>
<tr>
<td>_____ palpation</td>
<td></td>
</tr>
<tr>
<td>_____ auscultation</td>
<td></td>
</tr>
<tr>
<td>_____ inspection</td>
<td></td>
</tr>
<tr>
<td><strong>Multiple Choice Item:</strong> 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?</td>
<td></td>
</tr>
<tr>
<td>1. Second intercostal space at the right sternal border.</td>
<td></td>
</tr>
<tr>
<td>2. Third intercostal space at the left mid-clavicular line.</td>
<td></td>
</tr>
<tr>
<td>3. Fourth intercostal space at the left sternal border.</td>
<td></td>
</tr>
<tr>
<td>4. Fifth intercostals space at the left mid-clavicular line.</td>
<td></td>
</tr>
<tr>
<td><strong>Multiple Choice Item:</strong> 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?</td>
<td></td>
</tr>
<tr>
<td>1. 360 milliliters</td>
<td></td>
</tr>
<tr>
<td>2. 560 milliliters</td>
<td></td>
</tr>
<tr>
<td>3. 720 milliliters</td>
<td></td>
</tr>
<tr>
<td>4. 760 milliliters</td>
<td></td>
</tr>
<tr>
<td><strong>Fill in the Blank/Calculation:</strong> 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?</td>
<td></td>
</tr>
<tr>
<td>____ milliliters</td>
<td></td>
</tr>
</tbody>
</table>
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 examinee’s 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

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

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

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


Tired of press-release collections and superficial summaries? Subscribe to *Professional Licensing Report* for in-depth coverage of legal trends and regulatory policies in:

- entry standards
- credentialing
- scope of practice
- misconduct
- impairment
- technology
- interstate mobility, and more.

**Free sample issue:**

[www.plrnet.org](http://www.plrnet.org) or 206-364-1178

*Listed: Legal Newsletters in Print*
You're responsible for producing a dynamic testing program that delivers reliable, accurate results. That's why you should count on Prometric to bring you over 40 years of expertise in the testing field—including test development and delivery, data management and everything in between. So breathe easy. When you work with us, you'll ace your testing program.

For a free consultation and the latest schedule of webinars, call 1-856-PROMETRIC or visit our website at www.prometric.com/seminars.

© 2007 Prometric. All rights reserved. Over achieve™ is a service mark of Prometric.