NCSBN

N♠   *f Nursing*

accomplishments of an individual related to a specified course of study or occupational discipline (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). An important difference between licensure examinations and educational examinations is the stated purpose. Licensure examinations primarily focus on protection of the pubic, while educational examinations focus primarily on individual student progression. Because educational examinations are developed to assist with a variety of judgments related to curricular progression, educational achievement and academic ability, the stakes of the examination may range from relatively low stakes to high stakes depending on the criticality of the judgment, thus allowing varying degrees of rigor with respect to technical quality and validity (AERA, APA, & NCME, 1999). Regardless of the stakes of the examination, an important difference between licensure and educational examinations is the scope of the content included within the examination. Because educational examinations measure educational achievement or progression, content decisions generally reflect a distinct set of curricular objectives, while licensure examination content decisions generally reflect a much broader set of knowledge, skills and abilities (KSAs) necessary for competent practice as defined by an entry-level practice analysis.

Introduction to the NCLEX® Computerized Adaptive Testing (CAT) Experience

The NCLEX examination is different than a traditional pencil-and-paper examination. Typically, pencil-and-paper examinations administer the same items to every candidate, thus ensuring that the difficulty of the examination is the same across the board. Because the difficulty of the examination is constant, the percentage correct is the indicator of the candidate's ability. One disadvantage of this approach is that it is inefficient. It requires the high-ability candidates to answer all the easy items on the examination, which provides very little information about their ability. Another disadvantage is that guessing can artificially inflate the scores of low-ability candidates because they can answer these items correctly 25 percent of the time for reasons that have nothing to do with their ability. Instead, the NCLEX examination uses CAT to administer the items. CAT is able to produce test results that are more stable using fewer items by targeting items to the candidate's ability. The computer's goal during the NCLEX examination is to determine the ability of the candidate in relation to the passing standard. Every time the candidate answers an item, the computer re-estimates the candidate's ability. With each additional item answered, the ability estimate becomes more precise.

Each item that the candidate receives is selected from a large pool of items using three criteria:

1. The item is limited to a content area that will produce the best match to the test plan percentages. It ensures that each candidate's examination has enough questions from each content area to match the required test plan percentages.
2. An item is selected that the candidate is expected to find challenging. Based on the candidate's answers up to that point and the difficulty of those items, the computer estimates the candidate's ability and selects an item that the candidate should have a 50 percent chance of answering correctly. This way, the next item should not be too easy or too hard and the computer can get maximum information about the candidate's ability from the item.
3. Excludes any item that a repeat candidate has seen in the last year.

For CAT to work, the difficulty of each item must be known in advance. The degree of difficulty is determined by administering the items as pretest items to a large sample of NCLEX candidates. Because the difficulty of these pretest items is not known in advance, these items are not included when estimating the candidate's ability or making pass/fail decisions. When enough responses are collected, the pretest items are statistically analyzed and calibrated. If the pretest items meet the NCLEX statistical standards, they can be administered in future examinations as scored items.

The decision as to whether a candidate passes or fails the NCLEX-RN and NCLEX-PN examination is governed by three different scenarios:

Scenario #1: 95 Percent Confidence Interval Rule
This scenario is the most common for NCLEX examination candidates. The computer will stop administering items when it is 95 percent certain that the candidate's ability is either clearly above or clearly below the passing standard.

Scenario #2: Maximum-Length Examination Rule
Some candidate's ability levels will be very close to the passing standard. When this is the case, the computer continues to administer questions until the maximum number of items is reached. At this point, the computer disregards the 95 percent confidence rule and considers only the final ability estimate.

If the final ability estimate is above the passing standard, the candidate passes.

If the final ability estimate is at or below the passing standard, the candidate fails.

Scenario #3: Run-Out-of-Time (R.O.O.T.) Rule
If a candidate runs out of time before reaching the maximum number of items and the computer has not determined with 95 percent certainty whether the candidate has passed or failed, an alternate criteria is used.

If the candidate has not answered the minimum number of required items, the candidate automatically fails.

If at least the minimum number of required items were answered, the computer looks at the last 60 ability estimates:

If the last 60 ability estimates were consistently above the passing standard, the candidate passes.

If the candidate's ability estimate drops below the passing standard even once over the last 60 items, the examinee fails. This does not mean that the candidate must answer the last 60 items correctly. Each ability estimate is based upon all previous items answered.

Candidates may be administered multiple choice items, as well as items written in alternate formats. These formats may include but are not limited to multiple response, fill-in-the-blank calculation, ordered response and/or hot spots. All item types may include multimedia, such as charts, tables, graphics and sound.

Exhibiting the Core Values in Ensuring Test Validity

Because the NCLEX examination is a major component for entry into nurse practice, it is considered to be a high-stakes examination and therefore necessitates high standards to ensure the examination is valid. It must demonstrate both validity and reliability in order to allow BONs to make defensible licensure decisions. The validity of a licensure examination depends on both its ability to measure competencies necessary for safe and effective practice and to distinguish between candidates who possess these competencies from those who do not. The reliability of a licensure examination is its ability to yield consistent results; to pass or fail candidates possessing the same level of competency consistently. It is ultimately the core values of NCSBN that require the establishment of the validity of the examination and guide its maintenance over time. As demonstrated in the remaining sections of this article, the elements of collaboration, integrity, excellence and transparency are the building blocks upon which the most important aspect of the NCLEX examination, validity, is built and maintained.

Historically, validity has been defined as the degree of accuracy of a measure. A valid measure assesses all relevant aspects of a trait and only that trait. Messick (1989) states that the major concern of validity is not to explain any single isolated event or response to an item because these almost certainly reflect a confounding of multiple determinants; rather, the intent is to account for consistency in behaviors or item responses, which frequently reflects distinguishable determinants. (p. 14) Since the 1950s, psychometric research and reporting has referred to three major categories of validity: content-related, criterion-related and construct related.

More recently, the AERA, APA, and NCME have developed a unified concept of validity using an argument-based approach as noted in the *Standards for Educational and Psychological Testing* (1999):
> A sound validity argument integrates various strands of evidence into a coherent account of the degree to which existing evidence and theory support the intended interpretation of test scores for specific uses…Ultimately; the validity of an intended interpretation…relies on all the available evidence relevant to the technical quality of a testing system. This includes evidence of careful test construction, adequate score reliability, appropriate test administration and scoring, accurate score scaling, equating, and standard setting, and careful attention to fairness for all examinees. (p. 17)

As can be seen from this quotation, the *Standards for Educational and Psychological Testing* do not specify what type of evidence should be used to support that a licensure examination is measuring what is intended; rather, a validity argument is presented and evidence is accumulated to support intended interpretations and valid assumptions in major testing areas such as scoring, test administration, test content, standard setting and the interpretation/consequences of testing.

The *Standards for Educational and Psychological Testing* specify that licensure examinations should meet certain criteria which demonstrates their validity, as well as adhering to the intent of state and federal laws. NCSBN, as a test-producing entity, must be able to show that the NCLEX examination assesses candidates on the essential KSAs required to practice nursing upon entry into practice. A license is seen by the public as evidence that the practitioner possesses the necessary KSAs to provide safe and effective professional services; thus the licensing entity should ideally be independent from the professional and/or educational arm of the profession in order to provide independent validity evidence to the public (Clauser & Margollis, 2006). BONs meet this requirement of independence. Furthermore, licensing entities, such as BONs, are expected to establish requirements that are sufficiently extensive and demanding to meet the public's expectations. They accomplish this by permitting candidates to take the licensure examination after completing an approved nursing education program.

The validation process involves the development of an argument in support of the interpretations and inferences drawn from examination scores and evidence to support those arguments. "The interpretative argument can be represented as a chain or network of inferences leading from examination scores to conclusions to be drawn and decisions based on those conclusions" (Kane, 1992). The interpretative argument provides an explicit statement regarding the inferences and assumptions inherent in the argument and provides a framework for evaluating the proposed interpretation. All inferences and assumptions must be sound if the argument is to be considered valid. Moreover, an evaluation of plausible alternative interpretations should be considered as part of the validity argument (Kane, 2006).

The general validity assumption regarding the NCLEX examination is that the examination measures whether or not a candidate has the requisite KSAs needed for entry into nurse practice. This general validity assumption is supported by evidence scoring, generalizability, extrapolation and interpretation.

The primary validity argument of the NCLEX examination is provided by the rigorous processes in place for CAT. To ensure that the score key is reasonable, great care is taken to establish that the items being administered to candidates are fair and reasonable. Before each item is administered to nurse candidates, it goes through many reviews to ensure the item has only one correct answer and meets prescribed statistical criteria. Additionally, all items being administered to candidates are reviewed just prior to deployment of an item pool to make certain that the item keys are correct. All items in an item pool undergo a preliminary item analysis shortly after the pool is released in order to confirm that the item is meeting statistical criteria and there is only one correct answer. Further statistical analyses, such as making sure that the data fit the Rasch model, are performed.

Additional scoring validity is provided by the use of standardized procedures and conditions using professional testing centers. For the NCLEX, prior to an item pool rotating into the field, a cross functional group of professional staff meet to check that the item pool is being deployed correctly and that the item selection algorithm is working correctly. Accurate application of the scoring rules of an NCLEX examination is ensured by scoring each candidate's examination twice prior to the release of results, as well as other quality control procedures. Evidence to support this assumption includes consistency in the administration of the examination through the use of professional test centers that ensure registration, administration and scoring of NCLEX examinations are standardized across the U.S.

A great deal of time and effort is spent guaranteeing that the examination is administered under standardized conditions. One of the compelling reasons for transitioning NCLEX to a CAT examination was the structured testing environment with close monitoring by professionals and up-to-date technology resulting in few variations in testing conditions. Any variations in testing conditions are documented for NCSBN and can be used to invalidate a score, and if necessary, to ensure the scoring rule is applied consistently for all candidates. These are just a sample of the data and quality control procedures that provide evidence to support this assumption and the interpretive argument that the rules used to score NCLEX candidates are appropriate and applied consistently and accurately.

The validity argument for generalizability is "reliability or generalizability studies as well as judgments about the representative of the sample of observations included in the test" (Kane, 2006, p. 25). For the NCLEX, decision consistency is used as an index of reliability. Decision consistency represents the proportion of pass/fail decisions that would remain the same if all candidates were to be re-tested with parallel examinations. Regarding the representativeness of the sample of observations (i.e., the items administered to a candidate) at least four times a year, subject

test taker scenario is not likely to happen because of the many security procedures in place. For the NCLEX, these security procedures involve strict identification procedures, including the candidate's authorization to test (ATT), fingerprinting, palm vein scanning, as well as monitoring at professional test centers. This particular threat to a validity interpretation seems minimal.

A more plausible security threat to our validity argument is if candidates have advanced knowledge of NCLEX items. To mitigate this security risk, NCSBN uses two security firms to search for possible security risks and item content on

The NCLEX examination development process is an exemplar of collaboration, excellence, integrity innovation, and transparency. The examination development process includes several rigorous steps requiring the participation of

Supporting the core values of integrity, excellence and transparency, NCSBN requires that each experimental and operational pool be examined for potential DIF. Every item occurring in an experimental or operational pool with at least 50 focal group candidate responses and at least 400 reference group candidate responses is included in the DIF analysis to ensure items do not contain gender or ethnicity bias. Items that have a statistically significant difference using a t-test with significance .0001 and a difference in magnitude between focal and reference groups 0.50 logits are identified as possessing potential DIF. These items are then forwarded to the NCLEX® DIF Review Panel for review. The NCLEX® DIF Review Panel consists of a minimum of five members, including at least one male and at least three of the six ethnic focal groups included in the DIF analysis. No less than one individual must have prior experience on a DIF review panel, and at least one individual must have a linguistic background. In addition, one member of the panel must be a registered nurse (RN) with a current license. The NCLEX® DIF Review Panel is charged with reviewing all items to determine if the item possesses true bias toward any ethnicity or gender. Items determined to possess true bias are referred to the NEC for final disposition.

Member Board Reviews

Member Board Reviews represent a collaborative effort between NCSBN and BONs to help ensure that only items meeting the nurse practice act of all BONs remain in the NCLEX operational item pool. Twice a year, each BON is permitted to review items in the current operational pool for the purpose of identifying any item that would not be supported by their specific nurse practice act. Items identified as problematic by the BON are forwarded to the NEC for final disposition.

Master Pool Review

Major evidence for validity and technical quality of the NCLEX examination are established by the examination development process because nurse practice evolves over time based on nursing research, client complexity and workforce issues; the validity and the technical quality of the items measuring nurse competence must be maintained over time. Therefore, in support of the integrity and excellence of the examination, every item in the NCLEX item bank is required to be reviewed every four years. This review is completed by a panel of nurses from the four nursing regions with demonstrated expertise as a nurse and prior experience as an item reviewer. During this review all items are reviewed for currency to practice, item construction flaws and currency of evidentiary validation.

Exhibiting the Core Values through the Standard Setting Process

The standard setting process is another example of the core values of NCSBN in action, supporting collaboration, evidence-based decisions and excellence in upholding integrity of the examination scoring process. The standard-setting process utilizes information from a variety of stakeholders in nursing and testing, including nurses, educators,

Introduction to the NCLEX® Administration Experience for the Nurse Candidate

NCLEX candidates begin the process of taking the NCLEX by registering for the examination with NCSBN's test service provider. At this time the candidate also needs to apply for licensure with the BON where they wish to practice in order

and timely contact with NCLEX candidates and educators, answering questions and providing the most important information they need to be successful, including policies and procedures they need in order keep them aware of the rules surrounding the important, high-stakes nature of NCLEX information. Additionally, NCSBN Examinations communicates with BONs on a daily basis to discuss and resolve administration issues (e.g., ADA accommodations). BONs are also provided with a complete communication package, including the annual NCLEX Conference, topic-specific webinars and publications like the *NCLEX® Examination Candidate Bulletin* and the *NCLEX® Member Board Manual.*

Conclusion

NCSBN's adheres to the core values of collaboration, excellence, innovation, integrity and transparency in all aspects of support of its mission. As demonstrated by this article, the elements of examination development and administration of the NCLEX examinations are exemplars of the core values of NCSBN in action. The fundamental validity arguments related to the NCLEX examinations are directly supported by the transparency of the item development process, the collaboration required in the development of activity statement through the practice analysis process, innovation through development of alternate item styles, and the integrity and excellence principles adhered to during the standard setting process. The rigorous administration and security procedures of the NCLEX examination are additional examples of the use of NCSBN core values to support regulatory excellence.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing.* Washington DC: American Psychological Association.

Angoff, W. H. (1984). *Scales, norms and equivalent scores.* Princeton, NJ: Educational Testing Service.

Beuk, C. H. (1984). *A method for reaching a compromise between absolute and relative standards in examinations. Journal of Educational Measurement, 11,*

Wendt, A. & Worcester, P. (2000). The NCLEX-DIF process. *Journal of Nursing Education, 39*(4), 185-187.

Wendt, A. & Kenny, L. (2007). Setting the passing standard for the national council licensure examination for registered nurses. *Nurse Educator, 32*(3), 104-108.

Wendt, A., Kenny, L., & Riley, M. (2009). Fairness and sensitivity review. *Nurse Educator, 34*(5), 228-231.

Wendt, A. & Harmes, C. (2009). Memorability of innovative items. *CLEAR Exam Review, 20*(1), 16-20.

Woo, A., Wendt, A., & Liu, W. (2009). Readability of licensure examinations. *CLEAR Exam Review, 20*(1), 21-23.

N