Ensuring Validity of NCLEX® With Differential Item Functioning Analysis

By Ada Woo, PhD, and Marijana Dragan, MS

The National Council Licensure Examination (NCLEX®) is used for entry-level licensure of registered nurses and licensed practical/vocational nurses. Because of the high-stakes nature of the NCLEX, great care is taken to ensure all administered tests are psychometrically sound, the content is valid, and the tests are legally defensible. To minimize bias in the NCLEX, the National Council of State Boards of Nursing employs differential item functioning (DIF) analyses, a DIF review panel to analyze items for language bias and content relevance, a language sensitivity review, and readability analyses.

Owned and developed by the National Council of State Boards of Nursing (NCSBN), the National Council Licensure Examination (NCLEX®) is used for entry-level licensure of registered nurses (RNs) and licensed practical/vocational nurses in all U.S. states, Guam, Northern Mariana Islands, Virgin Islands, and American Samoa. In 2010, 197,779 NCLEX-RN® and 82,521 NCLEX-PN® tests were administered.

Because of the high-stakes nature of the NCLEX, great care is taken to ensure all administered tests are psychometrically sound, the content is valid, and the tests are legally defensible. This process includes thorough analyses of potential biases such as those related to gender and ethnicity.

Detecting Bias

A test is biased if it is less valid in measuring a test construct in one group than it is in another (Shealy & Stout, 1993). A biased NCLEX would be a test that measures nursing knowledge in one group of examinees and some unrelated construct, such as reading comprehension, in another. One way to assess potential testing bias is to investigate bias at the item level, using differential item functioning (DIF) analyses.

DIF exists when item performance of two or more groups of examinees differs after their ability levels are held constant. If a test item functions differentially between two groups of examinees, performance on the item may depend in part on the examinees' group classification. DIF analyses compare item performance of two groups of examinees, adjusting for examinees' ability. In the language of DIF analyses, *focal group* is the group of interest, generally the minority group, and *reference group* is the group with whom the focal group is compared (Angoff, 1993).

Several statistical methods are available for DIF detection. One of the most widely used is the Mantel-Haenszel procedure (Mantel & Haenszel, 1959). With this procedure, examinees from the focal and reference groups are matched on abilities. The odds ratio of the focal group answering a particular item correctly is then compared with that of the reference group answering the same item correctly. The procedure essentially estimates a common odds ratio across matched categories of examinees to ascertain if DIF exists. This procedure was used for DIF analyses on the NCLEX until 2007 (Wendt & Worcester, 2000). In October 2007, the Rasch Separate Calibration t-test (Wright & Stone, 1979) was used to increase sensitivity to small sample sizes and to stay congruent with the item calibration method used on the NCLEX. This method compares the difference between the calibrations—that is, the difficulties—of an item for the focal and reference groups. Multiple empirical research studies have shown that the Rasch Separate Calibration t-test is conceptually comparable to the Mantel-Haenszel procedure (Schulz, Perlman, Rice, & Wright, 1996).

Differential Item Functioning Analyses

Data from all U.S.-educated candidates are included in the routine NCLEX DIF analyses. Candidates educated outside of the United States are excluded because potential differences in nursing education curricula across different countries may confound results of the analyses, making them impossible to interpret. Among all domestically educated nursing candidates, gender and ethnicity are the focuses of NCLEX DIF analyses. Because females and Whites make up the majority of the candidate population, they are used as reference groups. For DIF analyses on genders, the performance of males is compared to the performance of females. For DIF analyses on ethnicity, focal groups include African American, Hispanic, Asian Other, Asian Indian, Native American, and Pacific Islander. Item performance for each is independently compared to that of the reference White candidates to determine the existence and magnitude of DIF. Table 1 shows the gender and ethnicity distributions of the

TABLE 1

Gender and Ethnicity Distributions of 2010 U.S.-Educated NCLEX Candidates

Candidate Characteristics	NCLEX-PN		NCLEX-RN	
	#	%	#	%
Female*	67,380	86.14%	142,892	87.04%
Male	10,842	13.86%	21,283	12.96%
African American	17,674	23.84%	16,170	10.63%
Asian Indian	958	1.29%	1,612	1.06%
Asian Other	4,484	6.05%	7,293	4.80%
White*	37,592	50.70%	104,310	68.59%
Hispanic	6,780	9.14%	10,747	7.07%
Native American	510	0.69%	985	0.65%
Other	4,753	6.41%	8,996	5.92%
Pacific Islander	1,396	1.88%	1,956	1.29%

Note. * denotes the reference group.

22,008 candidates did not provide information regarding ethnicities; 5,827 candidates did not provide information on gender.

248,224 U.S.-educated candidates who took the NCLEX-RN and NCLEX-PN in 2010.

The NCSBN conducts DIF analyses semiannually, covering all the operational (scored) and pretest (unscored) items administered in the past 6 months. The DIF analysis samples consist of all U.S.-educated candidates who took the NCLEX-RN or NCLEX-PN in the period being analyzed. All items analyzed must meet minimum sample size requirements to ensure statistically stable and practically meaningful results. Based on internal research, the NCSBN's policy requires that an item have at least 50 focal candidate responses and at least 400 reference candidate responses for DIF analysis. If minimum sample size requirements are met, items undergo separate calibrations on focal and reference groups. These calibrations are then compared, using independent sample t-tests. Items are flagged for further review if group calibrations differ by 0.5 or more logits at a significance level of 0.0001. These criteria ensure that the DIF statistics for flagging items have sufficient statistical power and that calibration differences are not spurious.

Content Review

Items identified as showing statistically significant DIF are further reviewed for content bias before removal from the NCLEX item pool. This review is done because equating DIF with item bias is conceptually improper. Although generally an undesirable characteristic in test construction, DIF alone does not render an item invalid (Nunnally & Bernstein, 1994). In many instances, items that show differential item performance between two groups of examinees are content appropriate and valid. Historically, NCLEX items related to obstetrics and gynecology are often flagged for statistical DIF. Female candidates tend to outperform male candidates on these items. Based on the NCLEX triennial practice analyses, however, obstetrics and gynecology are within the scope of entry-level nursing practice and should be included in the examinations. Similarly, many items related to operating medical equipment favor male candidates. Although statistically significant for DIF, these items are appropriate for the NCLEX as long as they measure knowledge relevant to entry-level nursing practice.

To determine if the items flagged for potential DIF are actually biased, they are forwarded to a subject matter expert panel for further content and language review. The NCLEX DIF Review Panel has at least five members, including members of at least three ethnic focal groups and one male. At least one member of the DIF Review Panel must have a general background in linguistics, and one member must be a licensed RN. The panel meets twice a year and reviews all items flagged for statistical DIF in the previous 6 months. Panel members review each item for potential language bias and content relevance for entry-level nursing. All items identified by the DIF Review Panel as biased are forwarded to the NCLEX Examination Committee, an oversight group of Boards of Nursing staff, for final decisions. Items deemed content irrelevant are removed from operational use immediately. In a typical meeting, the panel reviews more than 500 NCLEX items over 2 days. Less than 2% of all items flagged for statistical DIF are removed for true content bias.

Sample Items Identified in NCLEX DIF Analyses

Below are two NCLEX items identified as content biased. The first one illustrates bias towards an ethnic focal group; the second item favors females. Both items have been removed from operational use and are no longer part of the NCLEX item inventory.

Sample Item 1

The nursing care plan for a 74-year-old resident of a long-term care facility includes actions to promote the quality and duration of the client's nighttime sleep. Which of the following behaviors, if exhibited by the client, would indicate an appropriate action?

1. The client does mild calisthenics 1 hour before bedtime.
2. The client takes walks in the halls primarily in the afternoon.*
3. The client takes naps from mid- to late afternoon.
4. The client drinks warm tea before bedtime.

Note. *identifies the intended correct answer.

This item was removed from the NCLEX item pool because of its potential bias towards Hispanics. According to a DIF review panelist, option 4, "The client drinks warm tea before bedtime," could be perceived as a correct answer by Hispanic

candidates because drinking warm tea before bedtime may be considered an appropriate behavior in Hispanic cultures.

Sample Item 2

The nurse is caring for a 9-year-old client with bronchial asthma who was admitted with pneumonia. The client is on bed rest. Which of the following would be most appropriate to offer the client?

1. Coloring book and crayons
2. A toy stethoscope and syringe with needle
3. Beads and thread for making jewelry*
4. A radio and telephone

Note. *identifies the intended correct answer

This item was removed because it potentially favors females over males. According to one DIF review panelist, male candidates may not believe that making jewelry is a desirable activity for a 9-year-old child, especially if the child is male.

Conclusion

The goal of the NCLEX is to classify candidates into two groups: those who have adequate knowledge, skills, and ability to practice entry-level nursing safely and effectively and those who do not. Analyses are conducted on all NCLEX items with sufficient respondent data to ensure that the examination measures only nursing-related content and not extraneous constructs, such as gender and ethnicity. As described, DIF is a key step in ensuring the validity of the NCLEX.

NCLEX items also undergo language sensitivity review and readability analyses to ensure the validity of the test. Before being used on the NCLEX, all pretest items are reviewed by the Sensitivity Review Panel for appropriateness of language. The Sensitivity Review Panel is a subject matter expert panel similar to the DIF Review Panel. Members include persons from diverse ethnic backgrounds, nurses, and persons who are not nurses. The sensitivity review complements the DIF analyses because it provides another layer of content review at a stage during which no item response data can influence the reviewers.

All NCLEX operational item pools undergo readability analyses to ensure that item-reading load does not become a barrier to successfully completing the NCLEX (Woo, Wendt, & Liu, 2009). Results of readability analyses confirm that the effort required to read NCLEX items is far less than the effort required to read most nursing textbooks, and thus does not pose a hurdle to qualified candidates.

As the nursing scope of practice widens and patient acuity increases, the importance of allowing only qualified candidates to enter the profession grows. The NCSBN is tasked by its member boards of nursing to develop a valid and reliable assessment tool for entry-level nursing licensure. The NCSBN's goal is to use the NCLEX as a first step of public protection by following strict test development guidelines that meet or exceed industry standards. Through the painstaking test development process, NCSBN and its member boards can be confident that the NCLEX is a valid measurement of entry-level nursing competency.

References

Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–24). Hillsdale, NJ: Lawrence Elrbaum Associates.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*(4), 719–748.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory.* New York, NY: McGraw-Hill.

Schulz, E. M., Perlman, C., Rice, W. K., & Wright, B. D. (1996). An empirical comparison of Rasch and Mantel-Haenszel procedures assessing differential item functioning. In G. Engelhard & M. Wilson (Eds.), *Objective measurement: Theory into practice* (Vol. 3, pp. 65–84). Norwood, NJ: Ablex.

Shealy, R. T., & Stout, W. F. (1993). An item response theory model for test bias and differential test functioning. In P. W. Holland, & H. Wainer (Eds.), *Differential item functioning* (pp. 197–239). Hillsdale, NJ: Lawrence Erlbaum Associates.

Wendt, A., & Worcester, P. (2000). The National Council Licensure Examinations/differential item functioning process. *Journal of Nursing Education, 39*(4), 185–187.

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

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement.* Chicago, IL: MESA Press.

Ada Woo, PhD, is Senior Psychometrician and **Marijana Dragan, MS,** is Statistician in the Examinations Department at the National Council of State Boards of Nursing.