Assessing the Impact of English as a Second Language Status on Licensure Examinations

Thomas R. O'Neill, NCSBN; Casey Marks, NCSBN; Weiwei Liu, NCSBN

"As the developer of the national nurse licensure examinations (known as the NCLEX Examinations), the National Council of State Boards of Nursing (NCSBN) occasionally hears claims, like most examination programs, that its exams are unfairly biased against ESL candidates."

Background

It is an increasingly common assertion that language proficiency is generally considered a significant issue to ensure successful participation in many workplace activities in the United States (Quimby, July 2001). When the assertion is specific to the ability to enter into a licensed profession, the arguments are naturally extended to the educational requirements and/or the licensing examination of the profession. Typically, these assertions are manifested through a request for an accommodation, such as, accepting an international education as equivalent to a domestic education, providing US education in other languages, permitting ESL candidates to have additional time for assignments, permitting ESL candidates extra time on exams, allowing ESL candidates to use a translation dictionary during exams, etc.

As the developer of the national nurse licensure examinations (known as the NCLEX Examinations), the National Council of State Boards of Nursing (NCSBN) occasionally hears claims, like most examination programs, that its exams are unfairly biased against ESL candidates. The typical claim is usually framed around an anecdotal account of someone who is allegedly a fine nursing student, but has failed the licensing examination. To investigate the issue of potential bias, two issues were considered. First, do ESL graduates, in fact, tend to perform worse on the exams than graduates who are proficient in English? Second, is the same construct of nursing ability at work for both ESL and English proficient graduates?

Identifying ESL Candidates

NCSBN does not have access to empirical assessments of English proficiency for NCLEX candidates; however, the NCLEX application form does ask the candidate about their primary language. There are four possible responses: English, English & another language, another language, and no response. Typically, only a small percentage of candidates indicate that their primary language is another language or English & another language. These classifications are self reported indicators that are requested but not required. There are no efforts made to verify the accuracy of their claim. These categories, such as they are, were the basis for establishing whether a candidate was an ESL candidate or not.

Do ESL Candidates Perform Worse On The Exam?

Is there a difference in pass rates between U.S. educated ESL candidates and non-ESL candidates? If there is a difference (lower), a question that one might ask is: "Is the lower pass rate attributable to educational issues (acquiring the knowledge and skills) or some type of distortion in the exam?" Although this study did not address educational problems related to language, "Question 2" does consider whether individual exam items are biased and if so whether the bias was in favor of or against ESL candidates and more importantly the impact of item bias.

Sample Specifications

To adequately answer the research questions, the sample of candidates was limited to include only US educated NCLEX examinees. The inclusion of internationally educated examinees would confound the effects of language with the effects of curriculum. When candidates are educated in other countries using a curriculum designed for the scope of nursing practice in that country, not the US, it isn't clear whether observed differences on the NCLEX are attributable to curriculum difference, scope of practice differences, or English proficiency differences. Also, to prevent failing candidates from being included multiple times, only first-time candidates were included.

Disparate Pass Rates

Using 2002, 2003 & 2004 data (Tables 1 & 2), NCLEX-RN® and NCLEX-PN® pass rates were computed by the candidate's self-reported primary language status. The highest pass rates were for those examinees that indicated that English was their primary language or did not identify their primary language category. Pass rates for candidates who indicated that another language was their primary language or that English & another language were their primary languages was typically 10%-15% lower.

Table 1.

NCLEX-RN Pass Rates and Volume for 1st-time, US Educated Examinees by Primary Language Category.

	2002		2003		2004	
	Pass %	**# Tested**	**Pass %**	**# Tested**	**Pass %**	**# Tested**
English only	87.8%	62,289	87.8%	66,462	86.0%	75,617
English & Other Language	74.3%	4,393	76.0%	3,714	76.3%	3,898
Other Language	75.3%	1,431	76.3%	1,328	77.1%	1,681
Missing/Did Not Answer	85.8%	2,584	86.9%	5,227	84.6%	5,985
Total	86.7%	70,697	87.0%	76,731	85.3%	87,181

-Total also includes those first-time US educated candidates that did not indicate the category in which they belong.

Table 2.

NCLEX-PN Pass Rates and Volume for 1st-time, US Educated Examinees by Primary Language Category.

	2002		2003		2004	
	Pass %	**# Tested**	**Pass %**	**# Tested**	**Pass %**	**# Tested**
English only	88.4%	33,585	89.7%	37,990	90.8%	42,305
English & Other Language	70.1%	3,210	72.7%	3,062	75.7%	3.351
Other Language	72.1%	861	76.2%	807	76.9%	901
Missing/Did Not Answer	85.1%	740	88.2%	2,221	87.5%	2,736
Total	86.4%	38,396	88.2%	44,080	89.4%	49,293

-Total also includes those first-time US educated candidates that did not indicate the category in which they belong.

Is The Same Ability Construct At Work For Both ESL and English Proficient Graduates?

To answer whether the construct of nursing ability is generally the same for ESL and non-ESL candidates, it is helpful to have a framework to answer the question. The content of the NCLEX examinations are limited in scope by the test plan specifications. These specifications are derived from an extensive incumbent-based practice analysis. Furthermore, Rasch's (1960) model for dichotomies is used to estimate the difficulty of the items and the ability of the candidates. A consequence of using the Rasch model is that a hierarchy of items emerges and this hierarchy defines the construct, which is in this case "nursing ability." When the difficulty of an item and the ability level of a person is known, one can compute the probability of the person correctly answering the question; however this assumes that the hierarchy is invariant across subpopulations. When the hierarchy of item difficulties is dependant upon which subpopulation is responding, the construct is then not the same across these subpopulations. Under these conditions, the meaning of a correct response is different across groups. Differential Item Functioning (DIF) studies are used to check for these types of problems.

Differential Item Functioning

Differential Item Functioning is a method used to detect whether there is a difference in the probability of correctly answering a question across two groups of examinees after the ability of the two groups has been matched or controlled. This permits item-level bias to be detected. The procedure employed here compares calibrations based on the English only group with the calibrations based upon the ESL group. Using the standard errors for each pair of calibrations, a joint standard error was computed which was used to determine if the two calibrations were significantly different (Luppescu, 1991). The test was run with and without corrections for the accumulation of Type 1 error (alpha error).

Combining Groups

When performing DIF analyses, the sample size is important. When the number of responses per item is small, only very large bias effects can be detected. When the number of responses is large, then smaller bias effects can be detected. Given the number of US educated candidates who reported that their primary language was Another language or English and another language, it seemed useful to combine these groups into a "generic ESL" category. Also given that the pass rates for the English and another language group was below the pass rate for the Another language group, it seemed reasonable that this group might claim that they are also disadvantaged by language. The increase in statistical power attributable to the increased sample size seemed to outweigh the potential decrease in homogeneity of the ESL sample because it would permit more test items to be considered and the items could be calibrated with greater precision.

Sample Specification

The data selected for analysis were the responses from first-time, US-educated candidates taking the examination between April 1 and September 30, 2004. This sample was selected because it reflected a single item pool for each test (RN and PN) and contains a higher volume of examinees than the October – March time period. Combining language groups did help to boost the samples to sizes adequate to detect differences. Items for which there were fewer than 20 responses were excluded from the analyses. As a result, 76 RN and 54 PN items could not be analyzed. Of the 2000 items in the RN operational pool, 1924 were analyzed. Of the 1700 items in the PN operational pool, 1646 were analyzed.

DIF Results

The results, presented in Table 3, without the correction for the Type 1 error show no difference in the probability of a correct response for most (82-83%) of the items. The items that did show a difference were evenly split between providing an advantage for English speaking candidates (8% RN, 9% PN) and ESL candidates (8% RN, 9% PN). After the Bonferroni correction was used, only a trivial number of items continue to show a difference between groups and would not contribute to pass rate differences between groups.

Table 3.

Detection of DIF using an Item Recalibration Strategy on 1st-time, US Educated Examinees by Primary Language Category.

	RN	PN
Operational Pool	2,000	1,700
Excluded for Sample Size	76	54
Analyzed	1,924	1,646
Without Correction for Type 1 Error		
No Difference	1,605 (83%)	1,343 (82%)
Advantage English	162 (8%)	152(9%)
Advantage ESL	157 (8%)	151 (9%)
Using Bonferroni Correction for Type 1 Error		
No Difference	1,901 (99%)	1,641 (100%)
Advantage English	13 (<1%)	4 (<1%)
Advantage ESL	10 (<1%)	1 (<1%)
Mean Difference in Calibration	-0.02 logits	-0.03 logits

Difference classifications are based upon a 95% confidence interval. Negative mean differences indicate an advantage for English speakers. Positive mean differences indicate an advantage for the ESL group. Near zero mean differences indicate that there is not a systematic bias.

Figure 1.
The Stability of NCLEX-RN Item Calibrations Across English only and ESL Subpopulations.

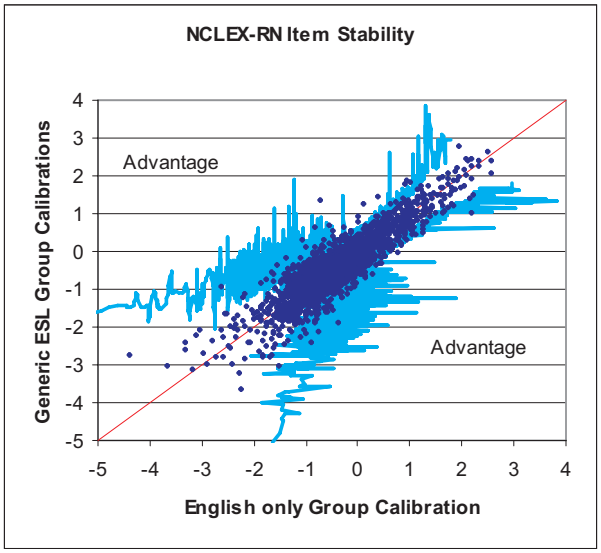
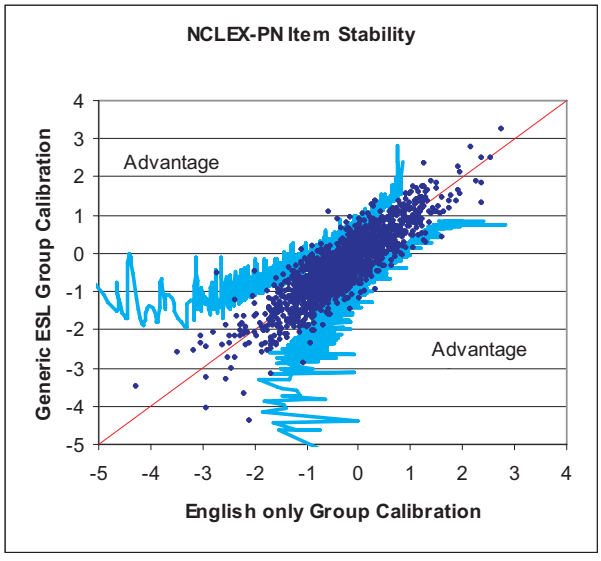
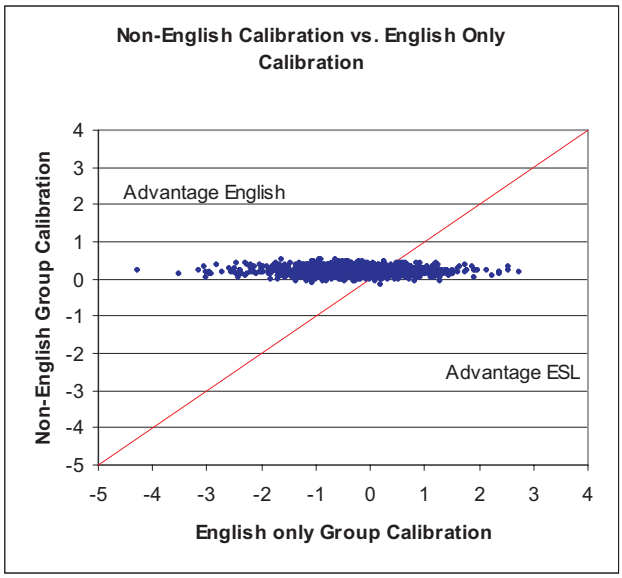


Figure 2.
The Stability of NCLEX-PN Item Calibrations Across English only and ESL Subpopulations.



The scatter plots indicate that generally the items calibrate the same way regardless of the language category the candidates' report. Most of the item calibrations fell near the identity line and within the 95% confidence interval. There were a few outliers, but they tended to be close to the 95% confidence interval and occurred in approximately equal numbers advantaging and disadvantaging the two groups. This comparability in calibrations indicates that the meaning of correctly answering a question is the same across groups. If the language demands of the items were governing the difficulty of the items then one would expect to see more deviation from the identity line. If the ESL population did not understand what was being asked at all, then the probability of a correct response would be approximately 0.25 for all items and the resulting calibrations would pile up around the same point on the Y-axis (example Figure 3). In this case, this clearly did not happen.

Figure 3.
What a Comparison of Item Calibrations Would Look Like if the ESL Group Spoke No English.



Discussion & Conclusions

As expected, the sum of these analyses indicates that there is a relationship between English proficiency and examination performance. Average pass rates clearly indicate that ESL candidates pass at a lower rate than English only candidates. These statistics, however, do not support the contention the bias resides in the examination. The DIF results indicate that the same construct of nursing ability is in effect across all language groups. Given that there is a disparate pass rate across groups, but the hierarchy of item difficulty is the same, one might hypothesize that lack of English proficiency may be an impediment to acquiring nursing knowledge and skills in US nursing programs.

NCSBN does believe that some degree of English profi-

ciency is necessary in order to practice even entry-level nursing in a safe and effective manner. For this reason, NCSBN has conducted studies to determine how much English proficiency is needed to be safe and effective (O'Neill, Tannenbaum, & Tiffen, 2005; O'Neill, Marks, & Wendt, 2005; O'Neill, 2004). Because the NCLEX was not designed to be an English proficiency test, it was written in such a manner that the level of English proficiency required to respond sensibly to NCLEX items was expected to be lower than the level of English proficiency required to safely practice entry-level nursing.

Readability

Because the purpose of the NCLEX examinations is to measure nursing ability, not reading ability, the reading demands of the test should not be so high that the readability of the text becomes a barrier to otherwise qualified candidates. Consequently, the difficulty of an item should be governed by the nursing content rather than the semantic or syntactic complexity of the text. To address this concern, NCSBN assesses the readability of each operational item pool before the pool is deployed for use. This is accomplished by evaluating three simulated tests from the new item pool: a minimum-length easy test, a maximum-length borderline difficulty test and a minimum-length difficult test. Because the items for these tests are from very different sections (with regard to item difficulty) of the item pool, it is unlikely that there would be overlapping items across the three tests. These items (approximately 18% of an operational pool) are then considered as a representative sample of the items in the operational pool. The samples are then analyzed using the Fry Readability Index (FRI) and the Lexile Framework® (Metametriks, October 2001). By policy, the readability level of the items should not be a significant barrier to passing for US educated examinees. More specifically, the readability level of an operational PN item pool should not exceed 1200 Lexiles and the readability level of an operational RN item pool should not exceed 1300 Lexiles. All operational NCLEX item pools are checked for compliance with readability policy before they are deployed.

Bias & Sensitivity

It should be noted that all NCLEX items are also evaluated for potential bias and sensitivity as part of the NCLEX item development process. The first evaluation of items for sensitivity takes place at NCLEX item writing and review panels. Then all items are evaluated by an independent panel of reviewers who are trained in the sensitivity review process prior to pretesting or any exposure to candidates. Any items that may be identified as unclear or insensitive at this juncture are forwarded to NCSBN's Examination Committee for further evaluation. After pretesting, items undergo a check for statistical item bias. After these analyses another independent panel of experts, who represent various ethnic and racial groups taking the NCLEX examinations, reviews any items that are identified as exhibiting statistical item bias. Items that this NCLEX-DIF panel identifies as exhibiting potential bias are referred to NCSBN's Examination Committee for final disposition. Despite these efforts, however, it must be noted that all well constructed examinations contain some items that exhibit some degree of bias.

Conclusions

Although all reasonable measures suggest that NCSBN's practices are in accordance with good testing practice, there are probably barriers that ESL candidates bring with them to the examination for which there are no reasonable remedies. This issue has been addressed earlier by the Standards for Educational and Psychological Testing (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 1999). The standards say that:

> Where effective job performance requires the ability to communicate in the language of the test, persons who do not have adequate proficiency in that language may perform poorly on the test, on the job, or both. In that case, the tests used for prediction of future job performance appropriately would be administered in the language of the job, as long as the language level needed for the test did not exceed the level needed to meet work requirements. (p. 91)

With regard to the meaning of answering particular items correctly, the results of this study imply that the same construct of nursing ability is in effect across both groups. Given that the hierarchy of item difficulty is the same across groups, yet there is a disparate pass rate, one might hypothesize that English proficiency may be a noticeable impediment to acquiring nursing knowledge and skills in US nursing programs.

This research was not able to identify any contributing factors beyond obvious issues of language competency that may impact performance on the NCLEX Examinations because candidates are not being negatively impacted by English language status. Although the results of this study were specific to two examinations, it is hoped that methods and processes described here will also be useful to other licensure examination programs that have substantial numbers of ESL candidates.