Version: 12/7/2006

Assessing the Unidimensionality of the NCLEX-RN

Thomas O'Neill and Michelle Reynolds
NCSBN, Inc.

The National Council of State Boards of Nursing, Inc. (NCSBN) is a not-for-profit organization that is composed of the jurisdictional boards of nursing in the United States and US territories. NCSBN's mission is to provide leadership to advance regulatory excellence for public protection. One of the many ways that NCSBN fulfills this mission is by providing its members (boards of nursing) with a defensible method of assessing a candidate's competence. Specifically, NCSBN creates and administers two minimal competency examinations, the National Council Licensure Examination for Registered Nurses® (NCLEX-RN®) and the National Council Licensure Examination for Practical Nurses® (NCLEX-PN®). All boards of nursing that are members of NCSBN use the NCLEX® as part of their licensing process.

Although adaptive tests provide many benefits, they also introduce many challenges. Sparse data is one of the major issues that is both a benefit and a challenge. It is wonderful that candidates may take shorter, less grueling tests because a small subset of the items is all that is needed, but the drawback is that the resulting data matrix is incomplete. In fact, when large item pools are used, the data matrices are quite sparse. With the introduction of the Rasch model (1960) and item response theory (IRT) in general, calibrating items and estimating person ability on incomplete sets of the data is no longer extraordinary. In fact the NCLEX examinations have used Rasch's (1960) model for dichotomous items since 1984 to calibrate test items and measure candidates' ability. Yet one of the requirements of measurement implied by the Rasch model is unidimensionality. Interactions between people and items that result in data that cannot be sufficiently order by a single continuum are multidimensional and degrade the measurement properties of the item calibrations and candidate scores. Therefore, it is important to periodically assess that the interaction of candidates and items is predominantly unidimensional. However, tests of dimensionality typically require complete data or near complete data designs. The methods available to assess multidimensionality in sparse data matrices seem relatively few. The two most popular are analysis of model-data fit and principle components analysis. This paper presents a method for testing the hypothesis of unidimensionality using PCA given a sparse data matrix and an example to illustrate it.

<u>Dimensions</u>

The NCLEX examinations were designed to measure a single construct, nursing ability. Nursing ability could have been conceived of as being composed of several separate constructs (client needs, nursing process, specialty area, etc.), but that approach would require the development of several different scales and passing criteria for each one. Instead, the more general, overarching construct of "nursing ability" which encompasses those more specific areas was selected because it was a more parsimonious model. This paper addresses whether a general construct of nursing ability is warranted given the observed data.

Before investigating whether test data manifests some degree of multidimensionality, it is important to have a clear understanding of what dimensions are and where they come from. A dimension is the imposition of a human organizational schema upon experience in such a way that it is coherent, useful, and represents a single continuum of more or less. Dimensions are not creations of nature ... the earth's surface rather than through the planet.17corend;4(t)4(h)-2(ae)4(The invention and use of geodesic distance does not invalidate the notion of straight-line distance. Geodesic distance merely represents a better theory of how sailors travel. The selection of a dimension and

METHOD

Unidimensionality is usually assessed by analysis at the form level. For written tests, the analysis is rather straightforward. The items on the test form are tested to see if they are measuring the same thing, often

Data

Two types of data were analyzed. The first was NCLEX-RN examination results collected from April 1, 2004 to September 30, 2004. During this period there were 89,116 examinees. The second data set was simulated to be comparable to the first data set with regard to the difficulty of the items available, the ability of the candidates testing, and the same rules for item selection and scoring. The simulated dataset was different from the real dataset in that the simulees' responses to the items were model to meet the

people of different ability encounter an item, the person with the higher ability ALWAYS has the higher probability of answering it correctly. Similarly, when a person encounters two items of different difficulty, the more difficult item ALWAYS has a lower probability of being answered correctly than the easier one. The philosophy behind Rasch's model is that there is a single continuum onto which both items and people are mapped. Because the items represent what the examinee can and cannot do, the ordering and relative spacing of the items articulates the construct. Subsequently, a person's ability estimate is then expressed as the point on that item continuum where the person has a 50-50 chance of correctly answering an item. It is immediately obvious that the invariance of the item hierarchy is crucial.

The dichotomous Rasch model specifies that the probability of a correct response is governed by the difference between the ability of the person, β_v and the difficulty of the item, δ_i. However, the difference $(\beta_v - \delta_i)$ can range from infinity to negative infinity, but the probability of a correct response is limited to the range of zero to one. Converting the probability to a log odds ratio solves the restriction of range problem. Expressed mathematically, the dichotomous Rasch model is specified as:

$$\left\{ \qquad \right\}$$

RESULTS

Scaling the Data Sets

Both the observed and simulated data were scored using Winsteps (Linacre, 2005). The distribution of item calibrations and person ability estimates for both data sets are illustrated in Figures 1 & 2. For ability estimates, the results across datasets were comparable, but not identical. Both datasets contained 89,116 examinees, but the average number of items administered [Observed data = 107.8, Simulated data = 97.8] was a little different (Table 1); however, there were no test records that contained fewer than 60 or more than 250 items. This indicates that the minimum and maximum limits imposed by the algorithm was working correctly. The ability estimates generated [Observed data, mean = 0.51 (0.80); Simulated data, mean = 0.60 (0.93)] were also similar, but not identical. The two datasets produced comparable person separation indices, although the index for the simulated data was slightly higher because the standard deviation of the simulated dataset was larger.

Table 1. Comparison of Ability Estimates for Observed and Simulated Data

Real Data

SUMMARY OF 89116 MEASURED Examinees VALID RESPONSES: 5.5%

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	55.2	107.8	.51	.23	1.00	-.1	1.01	-.1

Table 2. Comparison of Item Calibrations for Observed and Simulated Data

Real Data

SUMMARY OF 1973 MEASURED Items LACKING RESPONSES: 27 Items

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	2493.8	4867.3	.00	.04	1.00	-.5	1.01	.1
S.D.	1731.5	3350.2	.99	.02	.03	3.0	.06	3.1
MAX.	8953.0	12213.0	3.99	.14	1.26	9.9	1.58	9.9
MIN.	83.0	370.0	-3.31	.02	.78	-9.9	.71	-9.9

REAL RMSE	.05	ADJ.SD	.99	SEPARATION 21.15	Item	RELIABILITY 1.00
MODEL RMSE	.05	ADJ.SD	.99	SEPARATION 21.31	Item	RELIABILITY 1.00

S.E. OF Item MEAN = .02

Sim Data

SUMMARY OF 2000 MEASURED Items

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	2235.6	4356.2	.00	.04	1.00	-.6	1.01	.0
S.D.	1359.3	2731.8	1.02	.02	.02	1.9	.05	2.1
MAX.	5351.0	10089.0	3.43	.16	1.11	7.0	1.36	8.2
MIN.	158.0	278.0	-3.24	.02	.93	-5.3	.82	-4.6

REAL RMSE	.05	ADJ.SD	1.01	SEPARATION 21.78	Item	RELIABILITY 1.00
MODEL RMSE	.05	ADJ.SD	1.01	SEPARATION 21.88	Item	RELIABILITY 1.00

S.E. OF Item MEAN = .02

PCA of NCLEX residuals

A principle components analysis was performed on the standardized residuals from both datasets. Although data sets of this size can be calibrated and analyzed with regard to fit, displacement, and the like, rather quickly, PCA takes much longer. The results for the two datasets are summarized in Table 3. Although the datasets were not identical, the differences do not seem large enough to degrade the quality of the conclusions drawn. Across both datasets, the largest factor (factor 1) accounted for less than one fifth of one percent of the total residual variance. With a first factor that is this small, it is nearly impossible to argue that there is any noticeable structure at all.

This is good news for NCLEX, but it does not make for a good example to illustrate the need to identify the difference between a real first factor and an artifact of the data. He

Table 3. Comparison Summary of Observed and Simulated Data.		
	Observed Data	**Simulated Data**
Candidates	**89,116**	**89,116**
Items[1]	**1,973**	**2,000**
Total Residual Variance[1] **(in Eigenvalue units)**	**1,973**	**2,000**
Factor 1	**3.5 (0.18%)**	**1.4 (0.07%)**
Factor 2	**2.1 (0.11%)**	**1.4 (0.07%)**
Factor 3	**1.9 (0.10%)**	**1.4 (0.07%)**
Factor 4	**1.8 (0.09%)**	
Factor 5	**1.8 (0.09%)**	

Note: The largest factor (Factor 1) accounted for less than one fifth of one percent of the total variance in the residuals.

[1] In the observed data, 27 items were turned off and therefore not administered to any candidates. Ideally, the number of items and candidates in the simulated data should match the observed conditions exactly, but this minor difference should not substantially harm the interpretability of the results.

DISCUSSION

The Rasch model requires that there is a single dimension, only the difference between B_n and D_i matters. However, in combining several content areas into the general construct of nursing ability, there are concerns that there could be some multidimensionality. Rather than modeling it in a multidimensional model, the choice was made to hold it constant. That is the basis for our test plan specifications. As a result, we have controlled the multidimensionality to prevent vast difference from person to person.

The advantages of this method of testing for multidimensionality include simplicity in communication and the ability to accommodate sparse data matrices that are not missing at random. Methods that are sufficient and easy to communicate are important. A comparison of observed structure with ideal permits the less technically inclined reader to understand the comparison without having to be conversant in factor analysis.

The disadvantages of this method are primarily the laborious nature of adequately simulating the data and the amount of time that it takes to run PCA on a data matrix of this size. However, there are also some limitations that are attributable to the nature of the data. In an adaptive test, there are very few off-target items. As a result, no response is terribly unexpected, which makes it difficult to identify misfit to the model or multidimensionality. Therefore the conclusion that degree of multidimensionality is practically zero, should be treated somewhat skeptically. Although this sparse dataset does not indicate any multidimensionality, it is possible that a complete data matrix would. Similar investigations with untargeted pretest data could provide a test for how data missing at random would fall out.

Future enhancements could include other confirmatory methods to demonstrate that the factors make no difference. Also, if the same type of matrix will be routinely examined, it may be practical to run several

Figure 1. Observed Data

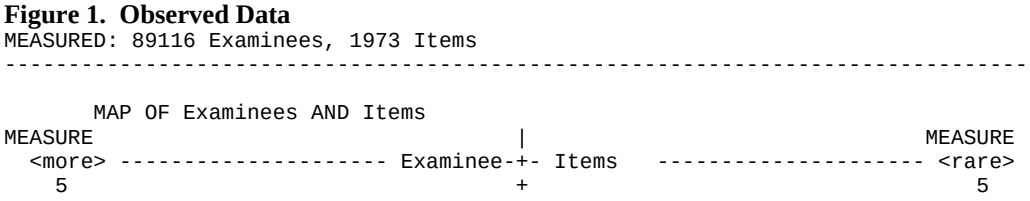

Figure 2. Simulated Data