Using Item Response Theory Models to Evaluate the Practice Environment Scale

Article *in* Journal of Nursing Measurement · August 2014

DOI: 10.1891/1061-3749.22.2.323

3 authors, including:

Dheeraj Raju
Cancer Treatment Centers of America
52 PUBLICATIONS **397** CITATIONS
SEE PROFILE

Patricia A Patrician
University of Alabama at Birmingham
97 PUBLICATIONS **1,559** CITATIONS
SEE PROFILE

JNM

Journal of Nursing Measurement

SPRINGER PUBLISHING COMPANY

www.springerpub.com/jnm

Journal of Nursing Measurement, Volume 22, Number 2, 2014

Using Item Response Theory Models to Evaluate the Practice Environment Scale

Dheeraj Raju, PhD
Xiaogang Su, PhD
Patricia A. Patrician, PhD, RN, FAAN

University of Alabama at Birmingham, School of Nursing

Background and Purpose: The purpose of this article is to introduce different types of item response theory models and to demonstrate their usefulness by evaluating the Practice Environment Scale. Methods: Item response theory models such as constrained and unconstrained graded response model, partial credit model, Rasch model, and one-parameter logistic model are demonstrated. The Akaike information criterion (AIC) and Bayesian information criterion (BIC) indices are used as model selection criterion. Results: The unconstrained graded response and partial credit models indicated the best fit for the data. Almost all items in the instrument performed well. Conclusions: Although most of the items strongly measure the construct, there are a few items that could be eliminated without substantially altering the instrument. The analysis revealed that the instrument may function differently when administered to different unit types.

Keywords: item response theory; Practice Environment Scale (PES); Rasch model; graded response model; rating scale model; partial credit model

The environment within which hospital nurses practice has long been recognized for its association with nurses' job satisfaction and turnover (Kramer & Hafner, 1989; McClure, Poulin, Sovie, & Wandelt, 1983) and more recently, for its contributions to patient outcomes. For example, more favorable work environments have been associated with lower patient mortality (Aiken, Clarke, Sloane, Lake, & Cheney, 2008; Kazanjian, Green, Wong, & Reid, 2005) and improved patient safety climate (Armstrong, Laschinger, & Wong, 2008). Aiken, Clark, and Sloane (2002) theorized that if nurses had adequate resources for patient care, the leadership support and authority to apply these resources to patient care, and good relationships with interprofessional colleagues that lead to enhanced teamwork, then hospitalized patients would receive higher quality care. Because these work environment attributes are associated with better patient and nurse outcomes, health care organization leaders should desire to improve the practice environment. To do this, they must have a suitable measure of this construct.

There are various instruments that measure the practice environment of hospital-based registered nurses (RNs), many of which have their roots in the early magnet hospital work of the 1980s and the subsequent development of the Nursing Work Index (NWI; Kramer & Hafner, 1989). The NWI captured the major facets of hospitals and units that were deemed good places to work because they attracted and retained a superior nursing workforce during a time of a serious nursing shortage. These early magnet hospitals attracted and

http://dx.doi.org/10.1891/1061-3749.22.2.323

retained nurses because they had superior practice environments characterized by nursing autonomy, authority, control over practice, and good working relationships with physicians (Kramer & Hafner, 1989). Empirical improvements to the NWI led to the development of the Revised NWI (NWI-R; Aiken & Patrician, 2000) and the Practice Environment Scale (PES; Lake, 2002). The NWI-R streamlined the response categories and devised conceptually formulated subscales: control, autonomy, and nurse–physician collaboration. The PES is a more parsimonious (31 items vs. 57) and contemporary rendition of the NWI-R. A confirmatory factor analysis of the PES yielded five subscales: nurse participation in hospital affairs; nursing foundations for quality of care; nurse manager ability, leadership, and support; staffing and resource adequacy; and collegial nurse–physician relationships (Lake, 2002). The PES asks nurses to what extent the 31 items characterize their current work environment. Nurses then respond on a 4-point Likert scale from 1 (*strongly disagree*) to 4 (*strongly agree*). Scores are averaged for each of the subscale scores, and a composite score is obtained by averaging all subscales to represent a global assessment of the practice environment.

The PES has been used in many countries outside the United States (Warshawsky & Havens, 2011) and has also been used in rather unique populations, for example, the military (Patrician, Shang, & Lake, 2010) and the Department of Veterans Affairs hospitals (Li et al., 2007). The PES has also been adopted by the National Quality Forum (2004) as one of its nursing-sensitive measures. Researchers have studied the instrument itself in comparison to other instruments (Bonneterre, Liaudy, Chatellier, Lang, & de Gaudemaris, 2008; Cummings, Hayduk, & Estabrooks, 2006) with mixed results. In many studies, the psychometrics of the PES were evaluated with standard methods that included reliability/consistency measures, exploratory and confirmatory factor analyses, and structural equation modeling (Gajewski, Boyle, Miller, Oberhelman, & Dunton, 2010; Hanrahan, 2007). To assess whether the PES is as parsimonious as it can be and to determine which particular items more strongly differentiate work environments, the authors applied item response theory models to and evaluation of the PES.

BACKGROUND AND CONCEPTUAL FRAMEWORK

Item Response Theory Models

Item response theory (IRT) models are successors of classical test theory (CTT; Lord & Novick, 1968; Spearman, 1904) that provide a predominant way of understanding and improving the reliability of psychological tests. An important limitation of CTT is that it considers the responses of a pool of respondents and ignores individual respondents. The IRT concentrates on item-level information and test-taker ability, as compared to CTT's focus on test-level information. The fundamental concept of IRT is that the probability of an examinee's expected response to an item question is the joint function of the examinee's ability and one or more parameters that characterize the question. The IRT approach considers the chance of getting any particular question right or wrong. Each item can be characterized by the probability of getting a right or wrong answer or receiving a high or low rating given the ability of the test taker (Kaplan & Saccuzzo, 2004).

For use with the PES, the IRT models facilitate an examination of the functioning or working performance of every item toward measuring the overall practice environment as rated by a nurse. Compared to other psychometric approaches, IRT models provide

dynamic and interactive diagnoses of each item and yield meaningful interpretations to assist in overall improvement in terms of the theoretical construct of the instrument. In IRT modeling, items that do not contribute to measuring the overall construct are known as "misfit" items. The outfit and infit statistics allow for the evaluation of misfit items and recommend suggestions. The suggested modifications might include actions such as removing some specific items or merging categories or levels of items, in efforts aiming for a better fit of the instrument to the construct and study population. In addition, IRT models can take into account the subscales used in designing an instrument and have features that allow one to compare the performance of items across different subpopulations via differential item functioning (DIF). For example, DIF can be useful in comparing the item performance of the PES between two groups, such as two unit types. The traditional survey instrument analysis includes assessing for reliability and validity using Cronbach's alpha and patterns of dependence among test items via factor analysis, respectively. IRT models analyze individual items in terms of their contributions to the overall construct measured by the instrument.

The IRT models emphasize that person and item parameters are distinguishable. The person parameter is often called "ability" (θ), which makes the best sense in answering binary (right/wrong) test questions. The ability to answer the test question (right/wrong) corresponds to the characteristic of the respondent, referred to as "trait." Because these traits are not measured directly, they are called latent traits. The latent trait for polytomous responses (i.e., Likert scales) defines the unidimensional construct that is explained by the covariance among all the item responses (Lord & Novick, 1968). Consequently, for polytomous responses, ability refers to a construct (i.e., attitude, anxiety, perception, etc.) that cannot be measured directly. In the PES, ability would refer to agreement that a particular item is present in the work environment. In addition to person ability, IRT models also take into account two important item parameters: item "difficulty" and item "discrimination." *Difficulty* is defined as the likelihood of getting a "correct" response (or high rating) for any particular item. Lower ability respondents are unlikely to answer difficult items and vice versa. Difficulty in terms of survey responses can be explained as the agreeability or the endorsement of a particular category of response on the item. Items with higher difficulty on a scale such as PES are those that tend not to be agreed with or endorsed. Discrimination illustrates the strength of an item's ability to distinguish respondents at different levels of the construct. Discrimination helps in separating the respondents in terms of ability (agreeability). An item with higher discrimination tends to better distinguish between higher ability and lower ability respondents over a smaller range of ability (agreeability), and therefore, a high-discriminating item is more reliable (DeMars, 2010). In other words, item discrimination indicates the relationship between the item and the construct. Higher discriminating items make it easier to distinguish between good practice environments and those that are poor or unfavorable.

Overall, the development and evaluation of the PES has been mainly based on factor analysis with focus on how individual question items are aligned along the subscales. IRT models provide a different viewpoint by checking the discriminating ability of each item, the appropriateness of the item scaling, and DIF. The purpose of this article is to identify the "best" fitting IRT models such as constrained and unconstrained graded response model, generalized partial credit model, Rasch model, and one-parameter logistic (1-PL) model for PES data collected for a nurse staffing and outcomes study (Patrician, Loan, McCarthy, Brosch, & Davey, 2010) and accordingly further explore the psychometric properties of the PES instrument. The remainder of the article is organized in the following

manner. First, CTT is reviewed, the basic IRT model will be explained, various types of models will be discussed, and finally, the application to the PES data will illustrate the use of this method of psychometric analysis.

Item Response Theory Models in Nursing Research

IRT models such as Rasch model, generalized partial credit model, and partial credit model have been extensively used, analyzed, revised, and modified instruments in nursing research (Fox, 1999; Hagquist, Bruce, & Gustavsson, 2009; Stump & Husman, 2012; Wang, Byers, & Velozo, 2008). These researchers concluded that IRT models can be very beneficial in nursing research. It is evident from literature that IRT models can be very beneficial in development, evaluation, and improvement of survey instruments in nursing research. However, most of the IRT papers published in nursing research journals used only one type of IRT model for analysis. Moreover, various IRT models have different assumptions about the nature of the instrument. For example, the 1-PL model is primarily concerned about item difficulty while assuming same discriminating ability of all items, whereas the two-parameter logistic (2-PL) model takes both discrimination and difficulty into consideration. Therefore, applying model selection criteria is critical. In current practice of Rasch analysis in nursing research, the authors rarely address the model selection issue. Therefore, this article introduces model selection criterion and different types of IRT models such as constrained and unconstrained graded response model, partial credit model, Rasch model, and 1-PL models to nursing research.

Item Response Theory Models for Dichotomous Responses

IRT models the response of a person on a test item as a function of some person characteristics and some item characteristics. In IRT, the item parameter estimation is sample-free, whereas respondents' characteristic estimation is item-independent. The standard mathematical model under IRT is represented as a logistic function; alternatively, other link functions such as probit (or normal ogive) or complementary log–log functions could be used as well. The original form of IRT was applied to dichotomous (yes = 1/no = 0) type responses, which correspond to right or wrong answers. Let Y_{ij} denote the response of *i*th person to the *j*th item. The general form of IRT model is given by its three-parameter logistic (3-PL) variant,

$$\Pr\{Y_{ij} = 1\} = c_j + \frac{1 - c_j}{1 + \exp\{-a_j(\theta_i - b_j)\}}, \tag{1}$$

where θ_i is the person ability or construct/latent trait parameter, a_j is the discrimination parameter for item j, b_j is the difficulty of endorsement for the *j*th item, and c_j is the so-called "guessing" parameter. We shall explain the meaning of these parameters step-by-step by gradually reducing the model. The guessing parameter, which supplies a lower asymptote of the probability as illustrated in Figure 1a, corresponds to the probability of getting the right answer "yes" to the *j*th item question by random guessing. Setting $c_j = 0$ in *Equation 1* leads to the 2-PL IRT model:

$$\Pr\{Y_{ij} = 1\} = \frac{1}{[1 + \exp\{-a_j(\theta_i - b_j)\}]}. \tag{2}$$

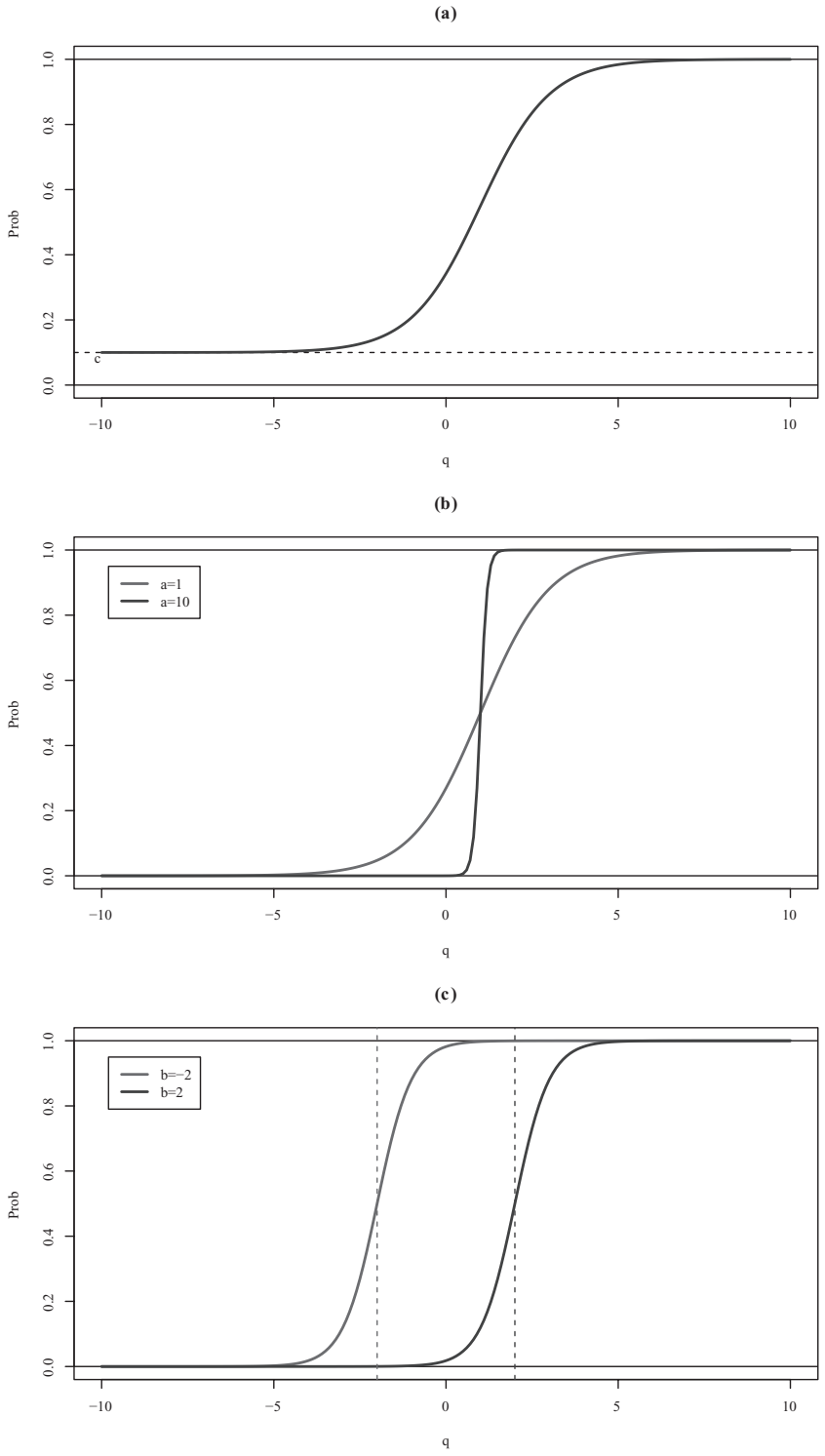

Figure 1. Illustration of the Rasch model: (a) nonzero guessing parameter c, (b) varying discrimination parameter a, and (c) varying difficulty parameter b.

Figure 1b plots the probability curve as a function of θ with two different choices of a. As it can be seen, the curve becomes steeper with a larger discrimination a and, as a goes to infinity, eventually converges to the threshold function that indicates (or discriminates) whether $\theta_i \geq b_j$ or $\theta_i < b_j$.

Figure 1b is made merely for illustrative purpose of demonstrating the effect of the discrimination parameter a. In real-world data, it is unlikely to see a highly discriminative pattern as shown by the blue curve. This could only possibly happen, for example, when a high school math question item is tested among either kindergarteners or college students.

Setting $c_j = 0$ and $a_j \equiv a$ in *Equation 1* leads to the 1-PL IRT model:

$$\Pr\{Y_{ij} = 1\} = \frac{1}{[1 + \exp\{-a(\theta_i - b_j)\}]}. \tag{3}$$

The 1-PL IRT model is often referred to as Rasch model; nevertheless, there is a small yet important difference between the 1-PL IRT model and Rasch model. De Ayala (2009) summarizes that Rasch and 1-PL models necessitate that items have a constant value of discrimination a but permit the items to differ in difficulty b. Their difference lies in the fact that the constant discrimination parameter a needs to be estimated in 1-PL IRT model, whereas it is set as 1 in the Rasch model. See also R package latent trait mode (ltm; Rizopoulos, 2006).

The discrimination a also can be treated as a model parameter. Figure 1c plots the probability as a function of ability θ with two different choices of difficulty b. It can be seen that difficulty b has to do with the location of the curve. The left curve corresponds to an easier item, whereas the right one corresponds to a more difficult item. With the easier item, a person with a lower ability value θ could also have a high chance of getting the right answer, but this is not the case for the more difficult item.

Ordinal Responses on a Likert Scale

Most instruments involve ordinal responses measured on a Likert scale. Suppose that each item contains K responses that range from 1 to K. To extend IRT models, both 1-PL and 2-PL IRT models are commonly used. Consider *Equation 2* for example, which can be rewritten as

$$\log\{\Pr(Y_{ij} = 1)/\Pr(Y_{ij} = 0)\} = a_j(\theta_i - b_j). \tag{4}$$

Note that the left-hand side of *Equation 4* is essentially the logarithm of the ratio of two probabilities. Various extensions of logistic regression models to ordinal responses differ in their ways of reformulating the two probabilities. In the most commonly used cumulative logit models (also called proportional odds models), the ratios of $\Pr(Y_{ij} \leq k)/\Pr(Y_{ij} > k)$ for $k = 1, \ldots, (k-1)$ are used. Other popular choices include continuation ratio models, where the ratios of $\Pr(Y_{ij} > k)/\Pr(Y_{ij} = k)$ for $k = 2, \ldots, K$ are considered, and adjacent categorical model, where the ratios of $\Pr(Y_{ij} = k+1)/\Pr(Y_{ij} = k)$ for $k = 2, \ldots, K$ are considered. When applied to IRT models, the graded response models follow a similar idea to cumulative logit models, whereas rating scale models, partial credit models, and generalized partial credit models are similar to the adjacent categorical models. No matter which model is used, the probability for the ith person to select the kth level of the jth item, that is, $\Pr(Y_{ij} = k)$, can be computed through the model equations. With that being said, different models place different constraints on the model parameters (i.e., person ability, item difficulty, and item discrimination) and hence carry different interpretations.

The same concepts in binary IRT models can be extended to a polytomous item. For example, in a hypothetical PES, the item (*Good working relationships with physicians*)

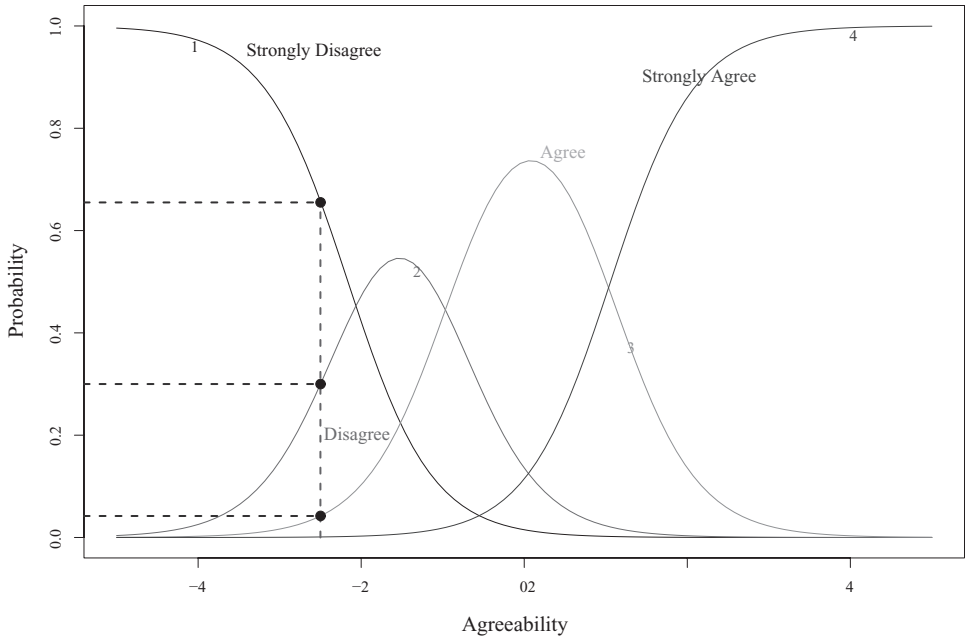

Figure 2. Example ICC for polytomous responses.

has four response categories (polytomous), and each response is shown by a curve (see Figure 2). The vertical line on Response 1 curve (*strongly disagree*) shows a respondent with agreeability ($\theta - 1.25$) with 1.25 standard deviations lower than the mean ($\theta = 0.00$). The perpendicular lines on the curves connecting the vertical line show probabilities of endorsing different categories of responses. In other words, a nurse with agreeability ($\theta = -1.25$) has a probability of 50% for endorsing *strongly disagree*, probability of 40% for endorsing *disagree*, probability of 10% for endorsing *agree*, and 0% probability of endorsing *strongly agree*.

Graded Response Model. In 1969, Samejima introduced the difference model also known as the graded response model (GRM) to analyze rating scales. There have been several modifications of GRM since its introduction. The GRM is appropriate for ordered polytomous responses such as Likert scales (Nering & Remo, 2010). The GRM for polytomous responses partitions the number of items m into $m - 1$ blocks. These blocks are often referred to as thresholds. The threshold can be defined as the level at which a likelihood of a response category below the threshold turns to a likelihood of success (Bond & Fox, 2001). A four-category Likert scale with scores ranging from $k = 1$ to $k = 4$ will have three threshold parameters (see Figure 3).

The GRM uses a two-step process to calculate the probability of a person's response on any item. The first step is to calculate the person's response that falls at or above a particular category. All possible probabilities to person response that fall at or above particular categories for Likert scale are (a) Response 1 versus 2, 3, and 4; (b) Responses 1 and 2 versus 3 and 4; and (c) Responses 1, 2, and 3 versus 4. The probability is expressed as

$$\log\left\{\frac{\Pr(Y_{ij} \geq k)}{\Pr(Y_{ij} < k)}\right\} = a_j(\theta_i - b_{jk}), \tag{5}$$

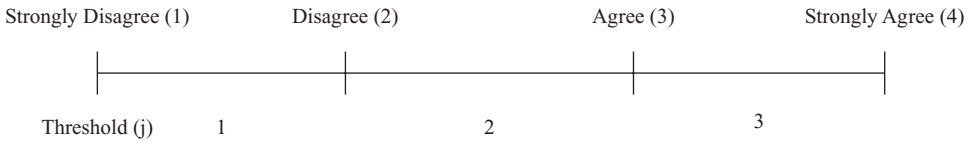

Figure 3. Likert scale illustrating thresholds.

for $i = 1,..., n$; $j = 1,..., p$; and $k = 1,..., (K - 1)$. Equivalently, $\Pr(Y_{ij} \geq k) = \frac{1}{[1 + \exp\{-a_j(\theta_i - b_{jk})\}]}$, where $\Pr(Y_{ij} \geq k)$ is the probability of responding k or higher category on item j, a_j is the discrimination of item j, b_{jk} represents the difficulty of endorsing the k or higher category on item j, and θ_i is person agreeability (Nering & Remo, 2010).

The second step involves calculating the actual response for each $k = 1$, 2, 3, and 4 categories. This is represented as follows:

$$\Pr(Y_{ij} = k) = \Pr(Y_{ij} \geq k - 1) - \Pr(Y_{ij} \geq k). \tag{6}$$

The model in *Equation 6* is referred to as unconstrained GRM. Setting $a_i \equiv a$ as constant is called as constrained GRM and can be considered equivalent to the Rasch model for ordinal data.

Partial Credit Model. Masters (1982) introduced partial credit model (PCM) for partial credit scoring known for polytomous items with ordered categories that can be applied for Likert scale analysis. The PCM similar to GRM partitions the number of items m into $m - 1$ blocks, and each of these blocks can be considered as modeling dichotomous responses. Considering the Likert scale with 1–4 as possible responses for any item, the PCM specifies the conditional probability of responding to any two pairs (Wu & Adams, 2007).

The general form of PCM is

$$\begin{aligned}\Pr(Y_{ij} = k + 1 | Y_{ij} = k \text{ or } Y_{ij} = k + 1) \\ = \frac{\Pr(Y_{ij} = k + 1)}{\Pr(Y_{ij} = k) + \Pr(Y_{ij} = k + 1)} = \frac{1}{[1 + \exp\{-(\theta_i - b_{jk})\}]},\end{aligned} \tag{7}$$

for $i = 1,..., n$; $j = 1,..., p$; and $k = 1,..., (K - 1)$. Consider the case $K = 4$ for example. In this case, the condition probabilities of responding are 1 or 2, 2 or 3, and 3 or 4. The probability of responding 1 or 2 is given by

$$\begin{aligned}P_{1/1,2} = (Y_{ij} = 1 | Y_{ij} = 1 \text{ or } Y_{ij} = 2) \\ = \frac{\Pr(Y_{ij} = 1)}{\Pr(Y_{ij} = 1) + \Pr(Y_{ij} = 2)} = \frac{1}{1 + \exp(\theta_i - b_{j1})},\end{aligned} \tag{8}$$

$$\begin{aligned}P_{2/1,2} = \Pr(Y_{ij} = 2 | Y_{ij} = 1 \text{ or } Y_{ij} = 2) \\ = \frac{\Pr(Y_{ij} = 2)}{\Pr(Y_{ij} = 1) + \Pr(Y_{ij} = 2)} = \frac{\exp(\theta_i - b_{j1})}{1 + \exp(\theta_i - b_{j1})},\end{aligned} \tag{9}$$

where θ_i is the latent variable, that is, ability of ith person and b_{j1} denotes the difficulty of endorsing jth item. *Equations 8* and *9* are in the form of dichotomous Rasch probabilities. Similarly, conditional probabilities for Responses 2 or 3 and 3 or 4 are calculated.

Generalized Partial Credit Model. The generalized partial credit model (GPCM) is the generalization of the PCM which allows the items within the scale to differ in slope parameter. Its general form is given by

$$\Pr(Y_{ij} = k + 1 | Y_{ij} = k \text{ or } Y_{ij} = k + 1) = \frac{1}{[1 + \exp\{-a_j(\theta_i - b_{jk})\}]}. \quad (10)$$

The only difference between GPCM and PCM is that additional discrimination parameter (a_i) is added to each item.

Rating Scale Model. Andrich (1978) proposed a rating scale model (RSM) in expanding the Rasch model to polytomous ordinal responses. The RSM is very similar to PCMs and can be viewed as a restricted PCM with additional assumption that the relative difficulty involved in the categories of each item should not vary from item to item. This consideration leads to the decomposition of the category difficulty parameter b_{jk} into two additive terms $b_{jk} = b_j + t_k$. The general form of RSM is given by

$$\Pr(Y_{ij} = k + 1 | Y_{ij} = k \text{ or } Y_{ij} = k + 1) = \frac{1}{[1 + \exp\{-(\theta_i - b_j - t_{k)}\}]}, \quad (11)$$

for $i = 1,..., n$; $j = 1,..., p$; and $k = 1,..., (K - 1)$. Thus, $(k - 1)$ location or threshold parameters t_k are used to characterize the relative locations of the K rating scale categories across all p items. Table 1 summarizes the characteristics of all IRT models.

Test and Item Information

The test and item information function extends the concepts of reliability, which traditionally is a single index that characterizes the average precision (or variability) of an instrument. However, IRT allows for nonuniform precision across the entire range of response scores. The name "information" comes from the statistical term "Fisher's information," which is directly involved in the variance–covariance matrix of the estimated parameters. In IRT models, the item information is obtained by expressing Fisher's information as a function of ability parameter θ. The information curve, which plots information versus θ, often looks mount-shaped. A highly discriminating item has tall, narrow information functions; they contribute greatly but over a narrow range. A less discriminating item provides less information but over a wider range. Thus, the area under the curve over a specified

TABLE 1. Item Response Theory Model Characteristics

Item Response Theory Model	Characteristics
One-parameter logistic	Discrimination equal and estimated; item difficulty varies
Rasch	Item discrimination = 1; item difficulty varies
Unconstrained graded response	Item discrimination and difficulty vary
Constrained graded response	Item discrimination constant; difficulty varies
Partial credit	Polytomous ordered items; discrimination and difficulty vary
Rating scale	Discrimination set constant; difficulty varies

range provides a good indicator for the importance or discriminating ability of the item to the construct. An item that has higher information has the capability to discriminate between individuals with different levels of agreeability (θ) on the item.

Model Selection and Feature Extraction

Different IRT models can be compared via model selection criteria such as Akaike information criterion (AIC) or Bayesian information criterion (BIC). Both AIC and BIC formulas for calculation use a term involving the parameters in the model and therefore is usually referred to as "penalized" model selection criteria. In calculating AIC indices, the set of models being compared does not contain the true model and true model is unknown, and although calculating BIC, the assumption is that the true model is included in the set of models being compared. Smaller values of indices indicate the best fitting model out of all the considered models (Kuha, 2004). It is worth noting that different software or packages might use different estimation methods to fit different Rasch models. For example, marginal likelihood and conditional likelihood are two common methods for estimating Rasch models in R program. However, the resultant log-likelihood scores cannot be directly compared with AIC. Similar to logistic regression, various features and diagnostic measures such as estimated coefficients, estimated probabilities, Pearson residuals, and goodness-of-fit tests can be extracted from a fitted model. These entities have been renamed using different terminologies in Rasch analysis and combined for further exploration and interpretations, such as infit/outfit statistics, option characteristic curve (OCC), item characteristic curve (ICC), and so forth. For example, ICC for each item plots all its category probabilities $\Pr(Y_{ij} = k)$ for $k = 1,..., K$ as functions of different values of person ability θ_i. RSMs would give us the same ICC for all items because of its model constraints.

Based on the standardized residuals, the (weighted) infit and (unweighted) outfit statistics provide similar ways to examine item or person fit. The unweighted outfit statistic indicates whether unexpected responses or outliers are found based on the person's ability for an item. Because the measure essentially involves sum of squared residuals, it is sensitive to outliers. The infit statistic is a weighted measure that down-weights outliers so that it could focus more on unexpected behavior that affects responses to items near the person's ability level. An infit statistic indicates the degree to which the observations for a particular item meet their model-based expectations. A higher infit and outfit statistic indicates a poor fit. In general, infit and outfit statistics perform similarly unless there are many outliers present. In principle, it is recommended that outfit be examined before infit statistic.

METHODS

Following human subjects approval at 13 military hospitals, the PES was administered to nurses in each study unit as part of a longitudinal multisite program of research on nurse staffing and adverse events (Patrician, Loan, et al., 2010). Nurses were surveyed with the PES annually over a 4-year period. Because it is possible that the same nurse could have answered the survey in multiple years, we use the data collected in the final year of the study only, where 936 nurses responded to the survey.

As part of data set preparation, a missing data analysis was conducted. Of the returned surveys, 215 (22.97%) had missing values or unanswered items. Items "active performance improvement program" and "preceptor program for newly hired RNs" had the highest

(5.45%, $n = 51$) missing values and item "good working relationships with physicians" had the least (1.92%, $n = 18$) missing. Although other methods such as imputation are available, the IRT models handle missing values naturally by assuming that these missing items were not administered to that person. In other words, IRT model outputs estimates only using available data.

Next, the data set was examined for lack of variability. Items and persons show no or little variation in responses, for example, a nurse (row) who has answered every item with a score of 3 or an item (column) that all nurses have rated 4. These rows and columns cannot be used for Rasch modeling because their lack of variability does little to help explain differences in practice environments.

Rasch Analysis Software

Some of the available Rasch analysis software are Winsteps, Rasch unidimensional measurement model (RUMM), Quest, ConQuest, Statistical Analysis System (SAS), and R. The R packages, ltm (Rizopoulos, 2006) and extended Rasch modeling (eRm; Mair & Hatzinger, 2007), were used to fit Rasch models for polytomous responses in this article. The package ltm uses approximate marginal maximum likelihood, whereas eRm package uses conditional maximum likelihood (CML) approach for estimation.

RESULTS

The final data comprised 888 records, with several nurses (rows) eliminated for lack of variability in responses but with no items removed. The overall proportion of *strongly disagree* responses ranged from 2% to 18%, *disagree* ranged from 6% to 33%, *agree* ranged from 34% to 55%, and *strongly agree* from 10% to 53%.

The overall means for the items ranged from 2.4 to 3.4, and the standard deviation ranged from 0.71 to 1.0. The item "High standards of nursing care are expected" had the highest mean (3.4), and the item "opportunity for staff nurses to participate in policy decisions" had the lowest (2.4) mean. The average age of respondents was around 37 years, with a standard deviation of 11.08 years. The overall instrument had a very good Cronbach's alpha score (.94), and all the five subscales had Cronbach's alphas of more than .80, indicating adequate internal consistency reliability (Pyrczak, 1999, p. 66). The Cronbach's alpha remained at least .93 even after excluding every item once from the instrument.

The IRT model unidimensionality assumption was evaluated using Kendall's rank-order correlation coefficient and chi-square test for pairwise association. The analyses indicated high correlation between items. Also, the first principal component (PC) of the item scores accounts for a greater portion of the variation in observed item responses. The proportion accounted for by PC1 is 37.84% as compared to the proportion accounted for by PC2, 6.14%, with a ratio of 6.165. These provided partial evidences for supporting unidimensionality. Conceptually, although five subscales were empirically derived from the PES, all the items make up the perception of the construct, practice environment. Besides, IRT models are moderately robust to departures from unidimensionality (Cooke & Michie, 1997).

Subsequently, we built IRT models such as constrained and unconstrained GRM, GPCM, Rasch model, 1-PL model, and RSM. First of all, we made efforts in identifying the best IRT model. We tried out several model choices as summarized in Table 2. The

TABLE 2. Item Response Theory Model Comparison

Model	AIC	BIC
Constrained graded response model	63380.87	63802.4
Unconstrained graded response model	**53453.17**	**54008.82**
Generalized partial credit model	53662.49	54218.14
Rasch model	53978.47	54395.21
1-PL model	53974.14	54395.67

Note. AIC = Akaike information criterion; BIC = Bayesian information criterion; 1-PL = one-parameter logistic.
Bold indicates the best IRT model.

unconstrained GRM and PCM had similar and low AIC and BIC indices when compared to other IRT models (see Table 2), indicating the best fit for the data. Therefore, both unconstrained GRM and PCM were further explored.

All the earlier mentioned IRT models can be categorized as generalized linear models (GLM), possibly with random effects modeling for the subject ability. GLM can be compared with AIC or BIC model selection criteria, which are aimed to balance off between model complexity (which varies dramatically with free or fixed discrimination parameters) and goodness of fit. A smaller AIC or BIC corresponds to a parsimonious model that provides a good fit to the observed PES data.

Figure 4 provides the item information curves for the unconstrained GRM. It can be seen that only few items have low peaks and nearly all of them contributes a considerable

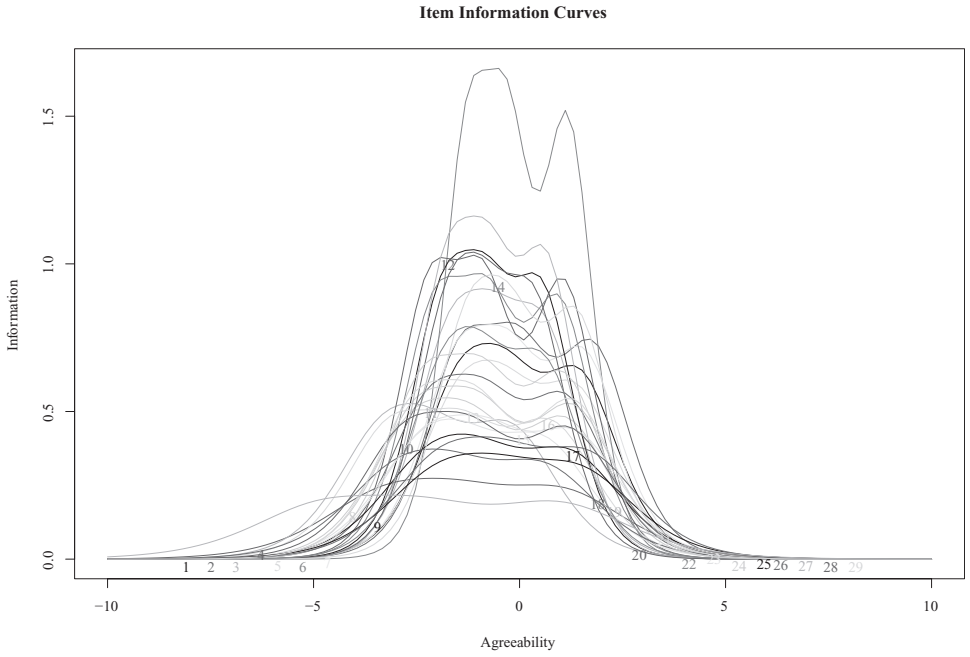

Figure 4. Item information of PES data.

amount of information. Items with high peak contribute significantly to the amount of information.

Items "administration that listens and responds to employee concerns" and "a clear philosophy of nursing that pervades the patient care environment" have the highest item information; that is, they are among the most important to the overall construct of practice environment (see Figure 4). The lowest information items were "good working relationships with physicians" and "chief nurse equal in power and authority to other top-level executives" (see Figure 4).

Figure 5 shows the item response category characteristic curves for the unconstrained GRM indicating that respondents with lower agreeability on most items have a higher probability of not endorsing (disagreeing) the item and respondents with higher agreeability for most items have a higher probability of endorsement (agreeing). Moreover, the curves from different items reasonably resemble each other, and the four categories are well-separated from each other, meaning that there is no need to merge categories.

R package "eRm" provides PCM plus features such as infit statistics, outfit statistics, and person-and-item map. The PCM was fit with 753 complete observations. The fit of the PCM of individual items was examined using the chi-square statistic that compares the observed responses with model-expected responses. Such an examination can also be made on persons. For any item that results in a significant chi-square statistic ($p < .001$), the item parameters are considered to be significantly different than those specified in the PCM (Muraki, 1997). In this sense, items "chief nurse equal in power and authority to other top-level executives," "preceptor program for newly hired RNs," "good working relationships with physicians," "patient care assignments that foster continuity of care,"

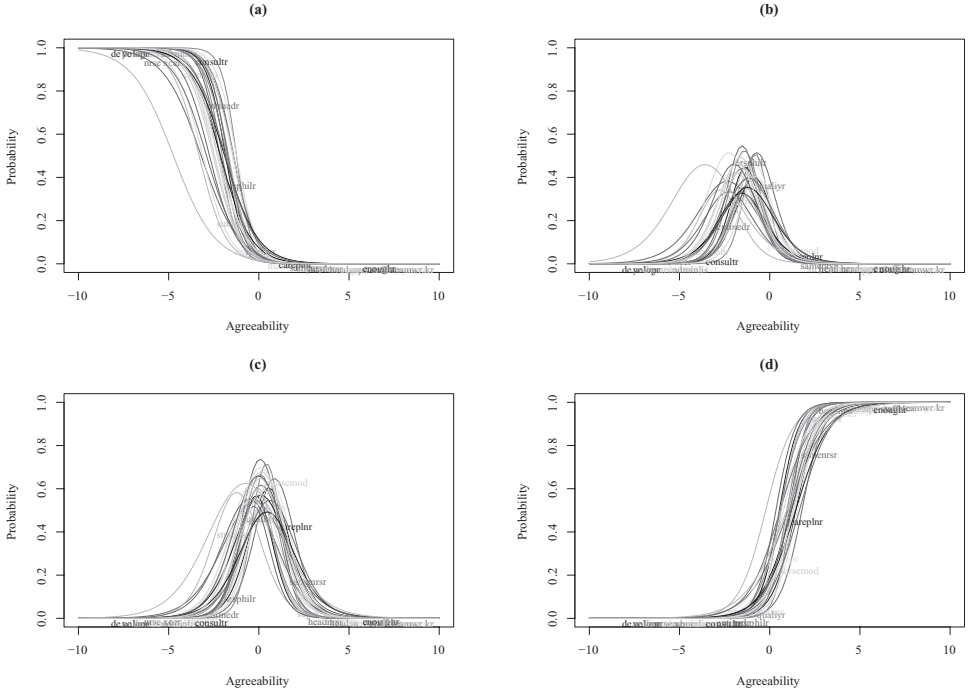

Figure 5. Item response category characteristic (ICC) curve for all four categories.

"enough registered nurses on staff to provide quality care," "enough staff to get work done," and "written up-to-date nursing care plans for all patients" had $p < .001$ and were therefore considered bad fit to the model (see Table 3). The same items (Table 3, boldfaced) had lower and similar discrimination parameter similar to the unconstrained GRM. Table 3 presents the chi-square, outfit, and infit statistics. It can be seen that same items with $p < .001$ showed higher infit and outfit statistics.

The DIF examines item similarity across different groups to identify differences in item performance between two groups. The entry "True" in DIF column indicates a large discrepancy difference in item performance by medical-surgical and step-down/critical care units (see Table 3). The 19 out of 29 items showed differences in performance between the two unit types. Furthermore, a PCM model was fit for RNs in medical-surgical units to compare the misfit items of the overall data. The model indicated that the same misfit items remain for RNs in medical-surgical units.

The item–person map describes the position of the item and threshold parameters along the construct (see Figure 6). The upper panel provides a histogram distribution of the nurse perception measure distribution, ranging from the lowest (left) to the highest (right) perception. Most nurses show positive perception toward their practice environments. The lower panel plots the location of for each PES item (solid dot) and thresholds for its categories. Note that the 1–4 levels have been recoded as 0, 1, 2, and 3 by the package. The items have been sorted from the most difficult (top) to the easiest (bottom). A difficult item (e.g., stndrdsr and dmsr) is the one on which few nurses rated high, whereas an "easy" item (e.g., policyr and praiser) is such that most nurses rated high or strongly agreed on this item. It can be seen that PES has reasonably good range coverage and is well-centered with respect to the person perception measure distribution. The last threshold of each item is so far out from location indicating that there have been a substantially high proportion of nurses who chose the highest rate (4) for every item.

DISCUSSION AND CONCLUSIONS

Overall, the PES performed very well regardless of the particular model used, with each item contributing to the overall construct. The test information function curve for all PES items was not severely skewed, which indicates that PES instrument is a reliable instrument for measuring the practice environment of nurses. Among the different IRT models that were compared, the unconstrained GRM and PCM had the lowest AIC and BIC scores, indicating the best fit for the data.

Using the GPCM to test for discriminating ability, several misfit items did not help to distinguish between good and bad practice environments. Therefore, they could be eliminated without substantially altering the construct. The lowest item information (lowest item discrimination) items were "good working relationships with physicians" and "chief nurse equal in power and authority to other top-level executives." Nurses working in military hospitals generally report very good nurse–physician working relationships, similar to nurses who work in magnet hospitals (Aiken, Clarke, Sloane, Sochalski, et al., 2001; Patrician, Shang, et al., 2010). This may be the result of the rank structure in the military that rewards individuals with a higher rank because of longevity and performance, not professional affiliation, thus leveling the nurse–physician hierarchy that is common in some civilian hospitals. Thus, in the absence of a professional hierarchy, nurses and physicians may develop more collaborative working relationships. Military chief nurses

TABLE 3. Rasch Analysis For Partial Credit Model

Item	Chi-Square	*p*	Outfit	Infit	Discrimination	DIF (Unit Type)
developr	733.463	.6790	−0.74	−1.02	1.570	False
policyr	679.233	.9730	−3.04	−3.33	1.688	False
dirmrsr	798.362	.1170	1.46	−2.23	1.754	True
nrsexcer	**1112.750**	**.0000**	**11.16**	**6.54**	**0.957**	**True**
advancer	767.291	.3410	0.55	0.31	1.501	True
adminlis	541.435	1.0000	−8.76	−9.32	2.431	True
nrsgovr	693.869	.9360	−2.36	−2.10	1.658	False
commitr	745.970	.5550	−0.23	0.34	1.364	True
consultr	636.457	.9990	−3.92	−4.55	1.893	False
cntinedr	725.442	.7500	−1.00	−1.51	1.465	False
stndrdsr	727.346	.7340	−0.60	1.03	1.326	False
nrsphilr	596.973	1.0000	−5.96	−6.40	1.932	False
nrscompr	712.931	.8430	−1.39	−0.88	1.346	True
qualiyr	612.336	1.0000	−5.40	−5.04	1.859	False
precepr	**904.359**	**.0000**	**4.49**	**2.64**	**1.256**	**True**
nursemod	732.393	.6890	−0.73	−0.98	1.428	True
careplnr	**853.218**	**.0060**	**3.46**	**3.50**	**1.187**	**False**
samenrsr	**934.578**	**.0000**	**5.40**	**3.47**	**1.113**	**True**
supervr	614.768	1.0000	−4.89	−6.31	2.001	True
headnrsr	668.650	.9870	−2.49	−4.17	1.859	True
praiser	655.512	.9950	−3.84	−3.90	1.806	False
headsupr	763.532	.3770	0.35	−1.73	1.617	True
supportr	790.284	.1620	1.40	1.67	1.292	True
problemr	658.648	.9940	−3.53	−3.17	1.564	True
enoughr	**987.249**	**.0000**	**8.11**	**5.59**	**1.090**	**True**
staffr	**859.674**	**.0040**	**4.00**	**3.61**	**1.179**	**True**
dmsr	**944.539**	**.0000**	**6.10**	**4.53**	**0.865**	**True**
teamwrkr	713.415	.8400	−1.40	−0.91	1.316	True
jntpracr	665.748	.9890	−3.16	−1.94	1.453	True

Note. DIF = differential item functioning.

do have equal authority status with the executive-level physician and administrator; all three leaders (nurse, physician, and administrator) report directly to the chief executive officer, who in military terminology is the hospital commander. This structure is standard in most military facilities, which could account for the poorer discriminating ability of this item.

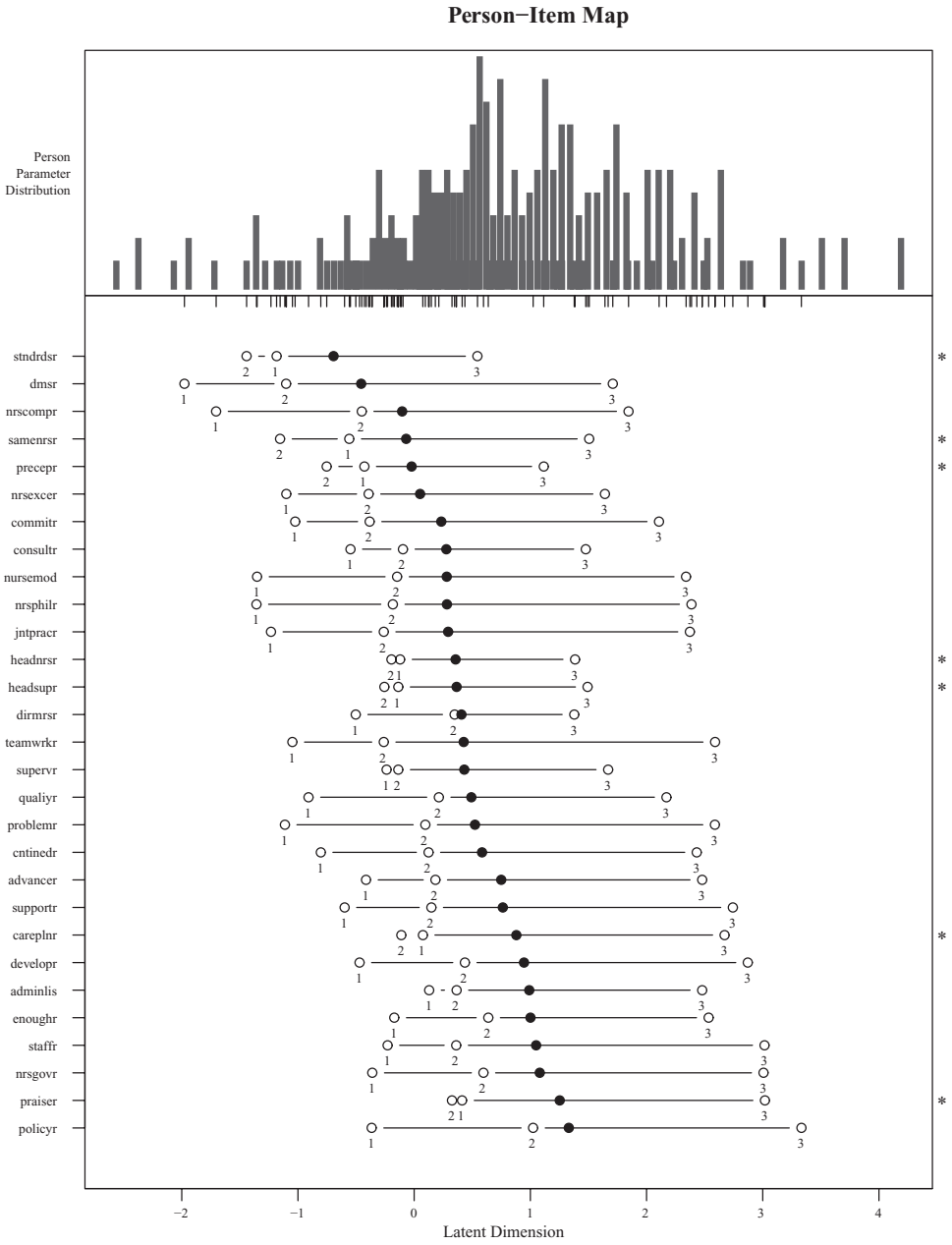

Figure 6. Person–item map for PCM.

The items with the highest discrimination, ability once again using the GPCM, were "administration that listens and responds to employee concerns" and "philosophy of nursing that pervades the patient care environment." Studies have demonstrated that nurses prefer environments in which they feel empowered and where contributions are valued (Kramer & Hafner, 1989), so these factors alone may very well distinguish good practice

environments from poor ones. The implication for leaders is that inculcating such characteristics in the organizational culture could enhance nurses' perceptions of their overall practice environment without the need to add additional resources.

The items "active performance improvement program" and "preceptor program for newly hired RNs" had the highest missing values suggesting that some nurses may not have understood the exact meaning of these items. The item "High standards of nursing care are expected" had the highest mean (3.4), and the item "opportunity for staff nurses to participate in policy decisions" had the lowest (2.4) mean. These items suggest that standards are perceived to be high perhaps because of the rank hierarchy in military hospitals and more rigorous policies such as dress codes. Also in this environment, many policies come from upper levels of military leadership even beyond the hospital, such as the surgeon general and/or nursing corps chief of each service, perhaps leaving nurses to feel disempowered at the unit level.

The items "chief nurse equal in power and authority to other top-level executives," "preceptor program for newly hired RNs," "good working relationships with physicians," "patient care assignments that foster continuity of care," "enough registered nurses on staff to provide quality care," "enough staff to get work done," and "written up-to-date nursing care plans for all patients" are considered poorly fit to the PCM, implying that they do not distinguish good versus bad practice environment and do not contribute in measuring the nurse practice environment. This again can be explained by the military-specific staffing models, which use a significant proportion of licensed practical nurses and unlicensed personnel who function at higher levels than seen in the civilian workforce. But it is surprising that there is no discriminating ability in the resources-related items, such as "enough registered nurses . . ." and "enough staff to get work done." Perhaps in military hospitals, expectations for numbers of staff are higher, leading to a general perception of adequate staffing.

The two most difficult (difficult to endorse, meaning most nurses rated these as the lower response categories) items are "High standards of nursing care are expected" and "good working relationships with physicians." Both items also have higher discrimination. The "easiest" (most nurses rated higher) items were "opportunity for staff nurses to participate in policy decisions" and "praise and recognition for a job well done."

As indicated by DIF, 19 out of 29 items showed differences in performance between the two unit types (medical-surgical and step-down/critical care). This means that the PES instrument may function differently when administered to medical-surgical and other unit types, suggesting the need to conduct separate analysis for surveys administered on different types of units. Some items need further examination and revision in order for PES to be applied in both unit types. Alternatively, separate PES instruments can be developed for different types of units. However, we do not feel these measures are necessary because we generally do conduct separate analyses of medical-surgical, critical care, and step-down units because of the differences in staffing patterns and patient acuity. This simply reinforces the need to do so. In addition, we also fit IRT model restricted to RNs in medical-surgical units only, and nearly the same set of misfit items were identified.

In summary, using IRT models was more informative than evaluating just the Cronbach's alpha (and alpha with each item removed) and an exploratory factor analysis (which has been conducted on the PES). Because IRT modeling allowed us to inspect the functioning of each item, we are able to evaluate which items have poor discriminating value in assessing the quality of the practice environment. Likewise, it gave us information about which items are most predictive of a good professional practice environment—items and their respective themes that have more actionable implications when a leader desires to improve the work environment of nurses.

The use of IRT models should be expanded in nursing research. They can provide additional information to further reduce unnecessary items in scales that do not add much discriminating value, thereby creating more parsimonious instruments. Furthermore, as we have seen with our analyses, several items seem to be more important in shaping the nurses' perceptions of the work environment, such as the item about administration listening to nurses' concerns. Leaders can and should use this actionable information to make positive changes to the work environment for nurses.

REFERENCES

Aiken, L. H., Clarke, S. P., & Sloane, D. M. (2002). Hospital staffing, organization, and quality of care: Cross-national findings. *International Journal for Quality in Health Care*, *14*(1) 5–13.

Aiken, L. H., Clarke, S. P., Sloane, D. M., Sochalski, J., Busse, R., Clarke, H., . . . Shamian, J. (2001). Nurses' reports on hospital care in five countries. *Health Affairs, 20*(3), 43–53.

Aiken, L. H., Clarke, S. P., Sloane, D. M., Lake, E. T., & Cheney, T. (2008). Effects of hospital care environment on patient mortality and nurse outcomes. *Journal of Nursing Administration*, *38*(5), 223–229.

Aiken, L. H., & Patrician, P. A. (2000). Measuring organizational traits of hospitals: The Revised Nursing Work Index. *Nursing Research*, *49*(3), 146–153.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573.

Armstrong, K., Laschinger, H., & Wong, C. (2009). Workplace empowerment and magnet hospital characteristics as predictors of patient safety climate. *Journal of Nursing Care and Quality*, *24*(1), 55–62.

Bond, T., & Fox, C. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York, NY: Lawrence Erlbaum Associates.

Bonneterre, V., Liaudy, S., Chatellier, G., Lang, T., & de Gaudemaris, R. (2008). Reliability, validity, and health issues arising from questionnaires used to measure Psychosocial and Organizational Work Factors (POWFs) among hospital nurses: A critical review. *Journal of Nursing Measurement*, *16*(3), 207–230.

Cooke, D. J., & Michie, C. (1997). An item response theory analysis of the Hare Psychopathy Checklist-Revised. *Psychological Assessment*, *9*(1), 3–14.

Cummings, G. G., Hayduk, L., & Estabrooks, C. A. (2006). Is the Nursing Work Index measuring up? Moving beyond estimating reliability to testing validity. *Nursing Research*, *55*(2), 82–93.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: The Guilford Press.

DeMars, C. (2010). *Item response theory*. New York, NY: Oxford University Press.

Fox, C. (1999). An introduction to the partial credit model for developing nursing assessments. *Journal of Nursing Education*, *38*(8), 340–346.

Gajewski, B. J., Boyle, D. K., Miller, P. A., Oberhelman, F., & Dunton, N. (2010). A multilevel confirmatory factor analysis of the Practice Environment Scale: A case study. *Nursing Research*, *59*(2), 147–153.

Hagquist, C., Bruce, M., & Gustavsson, P. (2009). Using the Rasch model in nursing research: An introduction and illustrative example. *International Journal of Nursing Studies*, *46*, 380–393.

Hanrahan, N. P. (2007). Measuring inpatient psychiatric environments: Psychometric properties of the Practice Environment Scale-Nursing Work Index (PES-NWI). *International Journal of Psychiatric Nursing Research*, *12*(3), 1521–1528.

Kaplan, R. M., & Saccuzzo, D. (2004). *Psychological testing: Principles, applications, and issues*. Belmont, CA: Wadsworth Publishing.

Kazanjian, A., Green, C., Wong, J., & Reid, R. (2005). Effect of the hospital environment on patient mortality: A systematic review. *Journal of Health Services Research & Policy*, *10*(2), 111–117.

Kramer, M., & Hafner, L. P. (1989). Shared values: Impact on staff nurse job satisfaction and perceived productivity. *Nursing Research*, *38*, 172–177.

Kuha, J. (2004). AIC and BIC: Comparisons and assumptions of performance. *Social Methods Research, 33,* 188–229.

Lake, E. T. (2002). Development of the Practice Environment Scale of the Nursing Work Index. *Research in Nursing and Health*, *25*(3), 176–188.

Li, Y. F., Lake, E. T., Sales, A. E., Sharp, N. D., Greiner, G. T., Lowy, E., . . . Sochalski, J. A. (2007). Measuring nurses' practice environments with the Revised Nursing Work Index: Evidence from registered nurses in the Veterans Health Administration. *Research in Nursing and Health*, *30*(1), 31–44.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Mair, P., & Hatzinger, R. (2007). Extended Rasch modeling: The eRm package for the application of IRT models in R. *Journal of Statistical Software*, *20*(9), 1–18.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

McClure, M., Poulin, M., Sovie, M., & Wandelt, M. (1983). *Magnet hospitals: Attraction retention of professional nurses*. Kansas City, MO: American Academy of Nursing.

Muraki, E. (1997). A generalized partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 153–164). New York, NY: Springer.

National Quality Forum. (2004). *National voluntary consensus standards for nursing-sensitive care: An initial performance measure set*. Washington, DC: Author.

Nering, M. L., & Remo, O. (2010). *Handbook of polytomous item response theory models*. New York, NY: Routledge Academic.

Patrician, P., Loan, L., McCarthy, M., Brosch, L., & Davey, K. S. (2010). Towards evidence-based management: Creating an informative database of nursing-sensitive indicators. *Journal of Nursing Scholarship*, *42*(4), 358–366.

Patrician, P., Shang, J., & Lake, E. (2010). Organizational determinants of work outcomes and quality care ratings among medical department registered nurses. *Research in Nursing and Health*, *33*(2), 99–110.

Pyrczak, K. (1999). *Evaluating research in academic journals: A practical guide to realist evaluation*. Los Angeles, CA: Pyrczak.

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory. *Journal of Statistical Software*, *17*(5), 1–25.

Spearman, C. (1904). "General intelligence," objectively determined and measured. *The American Journal of Psychology*, *15*(2), 201–292.

Stump, G., & Husman, J. (2012). The Nursing Student Self-Efficacy Scale: Development using item response theory. *Nursing Research*, *61*(3), 149–158.

Wang, Y. C., Byers, K. L., & Velozo, C. (2008). Rasch analysis of Minimum Data Set mandated in skilled nursing facilities. *Journal of Rehabilitation Research and Development*, *45*(9), 1385–1399.

Warshawsky, N. E., & Havens, D. (2011). Global use of the Practice Environment Scale of the Nursing Work Index. *Nursing Research*, *60*(1), 17–31.

Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement. A practical approach*. Melbourne, Australia: Educational Measurement Solutions.

Acknowledgments. This research is sponsored by the TriService Nursing Research Program, Uniformed Services University of the Health Sciences (Grant no. N10-C01); however, the information or content and conclusions do not necessarily represent the official position or policy of, nor should any official endorsement be inferred by, the TriService Nursing Research Program, Uniformed Services University of the Health Sciences, the Department of Defense, or the U.S. government. We also acknowledge Eileen Lake, PhD, RN, FAAN, School of Nursing, University of Pennsylvania, for her expert insights and editorial feedback. The authors declare no conflict of interest.

Correspondence regarding this article should be directed to Dheeraj Raju, PhD, The Center for Nursing Research, 1720 2nd Avenue South, NB 1020L, Birmingham, AL 35294. E-mail: seeth001@uab.edu