Nurse Education Today 35 (2015) 1181–1185

Contents lists available at ScienceDirect

Nurse Education Today

journal homepage: www.elsevier.com/nedt

The value of using test response data for content validity: An application of the bifactor-MIRT to a nursing knowledge test

Yuyang Cai *

Department of Chinese and Bilingual Studies, Hong Kong Polytechnic University, CF703, Hung Hom, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Accepted 16 May 2015

Keywords:
Bifactor multidimensional item response theory (bifactor-MIRT)
Content validity
Subject-matter experts (SME)
Test response data

SUMMARY

Aim: This paper aimed 1) to argue for the value of using test response data for content validation, and b) to demonstrate this practice using bifactor-multidimensional item response theory (bifactor-MIRT) for nurse education.
Method: The Nursing Knowledge Test (NKT) response data by 1491 nurse students from China were used for demonstration. Based on the content structure assumed by subject-matter experts (SME), a bifactor-MIRT model was constructed and tested. This involved five steps: dimensionality assessment, local dependence detection, model specification, calibrating and unit weighting.
Results: Dimensionality assessment results confirmed the content structure assumed by SME. Through local dependence detection and calibrating (i.e., item parameter check), items suspected of contaminating content were detected and those producing substantive harm were removed or constrained. Finally, content contributions by items to the overall scale and to their subscales were obtained through unit weighting.
Conclusion: Deficiencies residing in SME for content validation must raise attention. The study suggests the value of modeling test response data to compensate these deficiencies. The theoretical implication is discussed.

Introduction

Educational tests are commonly used to measure students' nurse knowledge (Redsell et al., 2003). To ensure measurement quality, it is fundamental to provide evidence of content validity (AERA, APA and NCME, 1999, 2014) that ensures "the content of the test is congruent with testing purposes" (Sireci and Faulkner-Bond, 2014, p. 101). Empirical endeavors to content validity usually rely on the judgment of subject-matter experts (SME) (Sireci and Faulkner-Bond, 2014), a practice also prevalent in nurse education (Beckstead, 2009). While this can help us understand certain facets of content validity, it risks unsatisfactory results with the test content due to complications resided in human judgment and other test procedures (Embretson, 1983; Sireci, 1998a). To overcome this limitation, empirical researchers turn to test response data (e.g., Colton, 1993; D'Agostino et al., 2011). This new orientation, however, has not been recognized in nurse education. The current study aimed a) to argue for the values of using test response data for content validity; b) to introduce bifactor-multidimensional item response theory (bifactor-MIRT) as an optional model for this practice, and c) to show how to apply this approach to evaluate the content validity of a nursing knowledge test.

Content Validity Evaluation

In a typical practice for content validity evaluation, a panel of SME are asked to link each test item with the test objectives, to assess the relevancy of the items to the content prescribed in the objectives, and finally, to judge if the items adequately represent the behaviors related to the intended content (Sireci and Faulkner-Bond, 2014; Waltz et al., 2010). This application, however, bears limitations. First, what it examines is human judgment per se rather than test content (Beckstead, 2009; Hogan, 2013). Assumption made in this way risks two types of confounding variances: uncertainty in the scale used for judgment data collection and uncertainty in human intuition. While many methods have been introduced to minimize the effect from the former (e.g., Lawshe, 1975; Newman et al., 2013; Penfield, 2003; Wilson et al., 2012), no progress has been made on the latter (Sireci, 1998b). The second limitation relates to information granurity. In reality, a test constructed under a single theory would consist of multiple content domains (Johnston et al., 2014). For test stakeholders such as teachers and students, a discrepancy between individual items as well as different domains in the reported score would have serious implications for diagnosing student performance (Leighton and Gierl, 2007). Making this differentiation, however, has proven to be difficulty for the SME (Murphy et al., 2013). In the paucity of studies that does deal with this difference (see Biddle, 2005; Haynes et al., 1995), SME are asked to rate directly the importance of different subcomponents. These ratings are then reflected in the test scores by balancing the number of items within each domain. The real contributions of different content

* Tel.: +852 3400 3824.
E-mail address: sailor_cai@hotmail.com.

http://dx.doi.org/10.1016/j.nedt.2015.05.014
0260-6917/

domains, however, are neither necessarily the same nor determined by the number of items (Rico et al., 2012). More objective approach is in need.

An idealistic solution would be to use test response data (Deville and Prometric, 1996). The value of test response data for content validity has been argued for decades ago. Lennon (1956) sees content validity as the interaction between test content and test responses. Ebel (1956) emphasizes that the only way to understand what content a test actually measures is to take the test by oneself. Guion (1977) asserts in his guidelines for content validity evaluation, "The response content must be reliably observed and evaluated" (p.7). This interactive view can find resonance among many other validity theorists (Embretson, 1983; Messick, 1989, 1995; Sireci and Geisinger, 1992). In short, whether test content is appropriate or not is one issue, whether it can actually activate behaviors related to the intended content is another. While SME judgment has been merited for understanding the first issue, test response data can be used to understand both, especially the latter.

Use of test response data can be found in a few educational studies. Colton (1993) used multivariate generalizability theory to evaluate the domain representation of test specifications of the ACT Mathematics Test (American College Testing, 1989). Deville and Prometric (1996) extended the multidimensional scaling method to model student's self-ratings of language competence. Ding and Hershberger (2002) applied structural equation modeling to examine the content meaning of each item and to testify whether the items measured the intended content domains at different levels. D'Agostino and his colleagues (2011) used confirmatory factor analysis with the 2004 Arizona state high school mathematics test. More recently, Schönbrodt and Gerstenberg's (2012) used exploratory factor analysis to examine the content clusters of motive inventories. Regretfully, no such exploration can be found in nurse educational research.

In nurse education, a test is usually designed to measure multiple domains of nursing knowledge. The test format is usually single multiple choice with four or more options and student responses are coded dichotomously. To provide granular information for content validity, an appropriate model is indispensable. The next section recommends bifactor-multidimensional item response theory (bifactor-MIRT) as an optimal method for our situation. We are aware that many other methods such as those applied in studies discussed above would suit our situation. However, a detailed discussion about the merits of those models falls out of the scope of this study.

Bifactor-MIRT

Recently, bifactor-MIRT has been valued as an idealistic method to evaluate test validity (Li and Rupp, 2011). This approach conceptualizes test multidimensionality as a set of uncorrelated factors: a general ability factor underlying all items and several domain-specific factors underlying different item subsets. Accordingly, the relationship between the probability of correct response to an item, given the general ability factor, its domain-specific factor and item characteristics, is formulated as:

$$P(y = 1|\theta_0,\ \theta_s) = c + \frac{1-c}{1 + \exp\{-[d + a_0\theta_0 + a_s\theta_s]\}},$$

where θ_0 is the general factor, θ_s is the domain-specific factor, c is the guessing parameter (lower asymptote), d is the item intercept, a_0 is the discrimination parameter on the general factor, and a_s is the discrimination parameter on its domain-specific factor. These item parameters can be estimated using computational methods such as Bock–Aitkin (Bock and Aitkin, 1981), Bifactor EM (Gibbons and Hedeker, 1992; Cai et al., 2011a, 2011b), Adaptive Quadrature (Schilling and Bock, 2005) and Metropolis–Hastings Robbins–Monro (Cai, 2010a, 2010b).

Bifactor-MIRT is deemed beneficial for our situation for several reasons. It can be used with dichotomous data in a confirmatory way (Reckase, 2009) to test the content structure assumed by the SME. Using this method, detecting contaminating content under the test becomes feasible. Moreover, it can be used to differentiate the relative content contributions of items to the overall scale or to their subscales. The former can be realized through item discrimination check after calibrating; the latter can be computed by extracting out the eigenvalues of the 2 (factors) by n (item discriminations) matrix for test items within the same subtest.

Evaluating the Content Validity of the NKT Using Test Response Data

Data Source

The current study used the Nursing Knowledge Test (NKT) response data. The NKT was an instrument used in a larger project that examined the relationship between nursing knowledge and nursing English reading ability. It was designed to measure knowledge in four subject areas: gynecology nursing, pediatrics nursing, basic nursing and medial nursing. Each subject comprised a subtest and tapped by six multiple choice questions. The test was constructed by two experienced healthcare teachers, who used to be professional nurses. Items came from the retired questions of the China Nurse Entry Test, a national licensing exam for Chinese nurses. Before test construction, they were informed of the purpose of study and particular content domains to cover. A sample question (in English) is:

Normally, an infant's anterior fontanel closes at:

A. 10 to 12 months B. two years old C. 18 to 20 months D. 12 to 18 months

Participants involved 1491 second-year nurse students (1465 females and 26 males) from eight medical institutions in China. They were all aged between 18 and 22 at the time of data collection. Before field entry, the author obtained ethical approval from his host university and had consent forms signed by the participating institution leaders and all participating students.

Procedures of Assessing the Content Validity of the NKT

The evaluation involved two phases: dimensionality specification and bifactor-MIRT modeling. Based on the SME, the test was specified to have five uncorrelated content dimensions: one general content domain (i.e., general nursing knowledge) representing content shared by all items and four particular content domains, one each representing the content exclusive to knowledge in gynecology nursing, pediatrics nursing, basic nursing, and medical nursing, respectively. The second phase comprised of five statistical steps: 1) assessing dimensionality; 2) detecting local dependence (LD); 3) model specifying, 4) calibrating and 5) unit weighting. Step 1 aimed to test the content structure assumed by the SME. Step 2 was to examine potential contaminating content at individual item or item cluster (two or more items) level; Step 3 was to determine the appropriate number of item parameters for best model estimation. Step 4 was to obtain item estimates to be used to identify the relative importance of individual items. These estimates were then used again in Step 5 to compute the relative importance of individual items to the overall scale and to their own subscales. Steps 1 to 4 were computed using the IRTPRO (Cai et al., 2011a) and Step 5 was computed by hand. The following section presents these results.

Results

Dimensionality Assessment

The results for the dimensionality assessment are presented in Table 1. The ΔG^2s due to successively adding four more domain-specific factors in the order of gynecology nursing, pediatrics nursing, basic nursing, and medical nursing to the general factor of nursing